2006

# A probabilistic approach for modeling and real-time filtering of freeway detector data

Shourie Kondagari
*Louisiana State University and Agricultural and Mechanical College*, skonda1@lsu.edu

# A PROBABILISTIC APPROACH FOR MODELING AND REAL-TIME FILTERING OF FREEWAY DETECTOR DATA

A Thesis

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Master of Science in Civil Engineering

in

The Department of Civil and Environmental Engineering

By
Shourie Kondagari
B.E., Osmania University, 2003
August 2006

# ACKNOWLEDGEMENTS

I wish to deeply acknowledge my advisor, Dr. Sherif Ishak, for providing me support financially and encouraging me through out the course of study. I would like to thank him for being patient with me during this research study and for shaping my career. I thank Dr. Chester Wilmot and Dr. Brian Wolshon for being on my defense committee, and for their suggestions and advice. Deep appreciation goes to my colleague, Ciprian Alecsandru, for helping me out with the much needed data set for this research study.

I am especially grateful to parents and my brothers (Rahul and Dheeraj) for supporting and motivating me throughout my life. Special thanks to all my friends (Ravi, Naveen, Sandeep, Kowndi, Santosh, Anil, Prashanth, Eshwar, Pradeep, Santosh Somireddy and Vamshi) who had made my stay and study at LSU, a most memorable one.

I dedicate this work to my parents Shri K.Mahendra Nath and Shri K.Pushpa Latha who constantly encouraged me to pursue higher studies and extended their love and support so far.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# ABSTRACT

Traffic surveillance systems are a key component for providing information on traffic conditions and supporting traffic management functions. A large amount of data is currently collected from inductive loop detector systems in the form of three macroscopic traffic parameters (speed, volume and occupancy). Such information is vital to the successful implementation of transportation data warehouses and decision support systems. The quality of data is, however, affected by erroneous observations that result from malfunctioning or mis-calibration of detectors. The open literature shows that little effort has been made to establish procedures for screening traffic observations in real-time. This study presents a probabilistic approach for modeling and real-time screening of freeway traffic data. The study proposes a simple methodology to capture the probabilistic and dynamic relationships between the three traffic parameters using historical data collected from the I-4 corridor in Orlando, Florida. The developed models are then used to identify the probability that each traffic observation is partially or fully invalid.

# 1. INTRODUCTION

## 1.1 General

Real time traffic information is vital to a variety of advanced operation and management functions undertaken by traffic management centers. The advent of new monitoring technologies has led to nationwide implementation of traffic surveillance systems on major urban freeway segments. Currently, several hundreds of freeway miles are instrumented with traffic surveillance devices such as electro-magnetic detectors, video detectors, radar detectors, and many others, all of which are primarily installed to improve the operation, safety, and productivity of our surface transportation network. These surveillance systems collect large amounts of real-time traffic data, sometimes on the order of a few gigabytes per day, and communicate them to traffic management centers (TMCs) to support critical functions such as incident detection, travel time and delay predictions, congestion management and other emergency services.

Advanced Traffic Information Systems (ATIS) need real time traffic data to disseminate real time traffic information to transportation system users via internet, in-vehicle navigation systems, variable message signs etc. Such information assists travelers in making better pre-trip planning and en-route decisions that affect their departure time, and choice of destination, mode, and route. Such decisions can effectively reduce travel costs in terms of travel time and delays. For transportation system providers, traffic information is essential for performance monitoring and decision support systems, which can be greatly influenced by the quality of traffic data. To date, robust data screening methods have not been fully developed to control the quality of data before its archiving, dissemination to the public, or use in relevant applications.

## 1.2 Problem Statement

Increasing need of the real-time traffic information to carry out advanced traffic operations and management functions emphasizes the importance of quality control of traffic data. Detection of erroneous observations also leads to identification of calibration problems with traffic monitoring devices and prompts for a quicker and more efficient maintenance service. Therefore, there is an urgent need for a robust real-time traffic data screening algorithm that is capable of identifying the probabilities of partially or fully invalid observations.

## 1.3 Objectives

This research study proposed probabilistic approach for real-time freeway traffic data screening. The proposed approach differs from the deterministic approach in that it does not explicitly confirm the validity of an observation but attempts to quantify the likelihood that such observation is valid. This research study considers the dynamic and probabilistic changes of traffic conditions to develop a real-time data filtering algorithm. The methodology focuses on building a family of models that capture the probabilistic variations of the three macroscopic parameters (speed, occupancy and volume count) collected from Inductive loop detectors. These probabilities provide the basis for validating observations using a user-specified confidence level.

The main goal of this research was to develop a real-time data screening algorithm by considering the stochastic variations in traffic conditions. This can be achieved by accomplishing the following objectives:

1. Develop a methodology to examine the probabilistic nature of the three macroscopic traffic parameters (speed, volume and occupancy), considering the stochastic as well as dynamic changes in the traffic conditions.
2. Model the probabilistic nature of the three traffic parameters and evaluate the calibrated model by performance measures.
3. Derive a data screening algorithm based upon the consistency of the probabilistic relationships as reflected by the model, and devise a strategy to further identify the partially valid observations (i.e. observations which have one or more invalid parameters).
4. Demonstrate using a sample data set how the newly developed data screening methodology can be applied to freeway traffic data in real time.

# 2. LITERATURE REVIEW

## 2.1 Introduction

Traffic surveillance systems are primarily used to monitor and collect traffic information from urban freeways. Traffic monitoring equipment can be classified as road-based and vehicle-based. Loop detectors, Closed Circuit Television (CCTV), sensors etc, are examples of road-based detection systems. Vehicle-based traffic surveillance systems include probe vehicles that are equipped with tracking devices, such as transponders, to track the location of vehicles over time. Figure 1 depicts how traffic information is relayed to TMCs from different monitoring equipment.

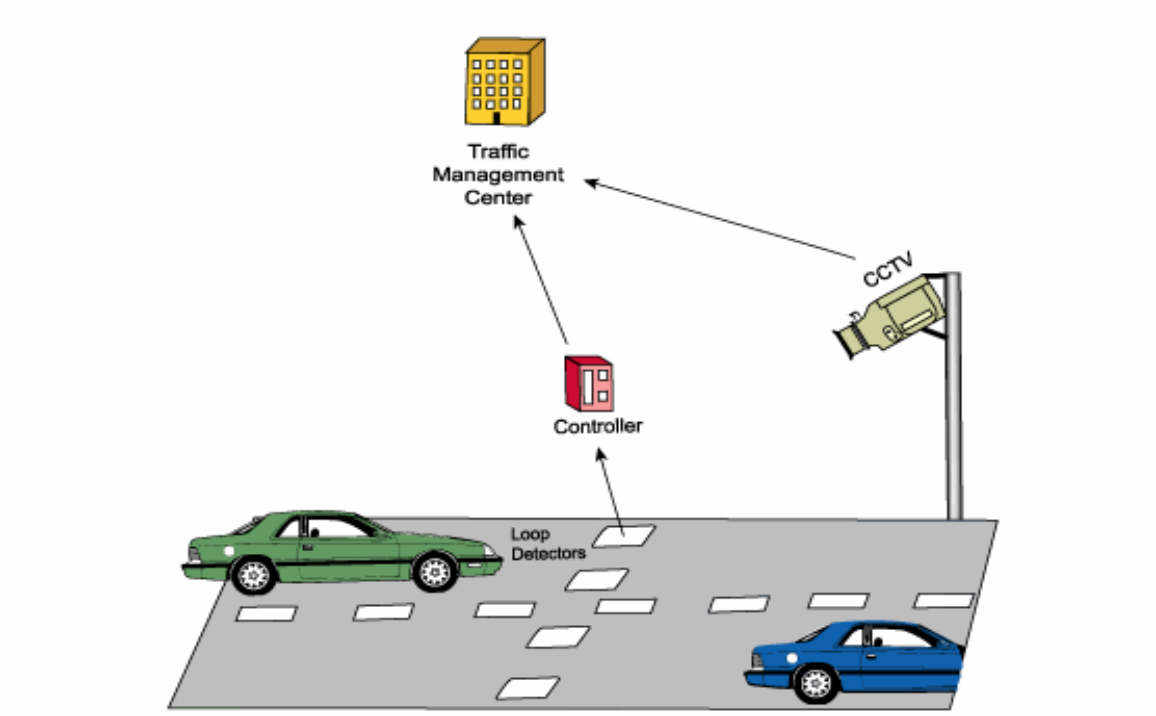


Figure 1: Traffic Surveillance System

The following section presents information about loop detectors, which constitute majority of real-time traffic monitoring devices.

## 2.2 Loop Detectors

Inductive loop detectors remain to be the most commonly used device for freeway surveillance and incident detection systems. Inductive loop detectors are constructed by cutting a slot in pavement and placing one or more turns of wire in the slot (see Klein, 2001). The wire is then covered by a sealant. The size of the loop detector ranges from 6-ft x 6-ft (for normal loops) to 6- x 40- to 70-ft (for long rectangular loops). Loops detectors collect vehicle count, lane occupancy and vehicle speed at intervals of 20 to 30 seconds and relay such information to traffic management centers.

Loop detectors operate on the 'principle of inductance'. Inductance is generated in a loop circuit due to current passing through a loop detector coil buried in the pavement. Loop detectors consist of four parts: a wire loop of one or more turns of wire embedded in the roadway pavement, a lead-in wire running from the wire loop to a pull box, a lead-in cable connecting the lead-in wire at the pull box to the controller, and an electronics unit housed in the controller cabinet (see Klein, 2001). When a vehicle passes over a loop detector it causes change in the initial inductance and the pulse is transmitted to the controller placed at the side of the pavement indicating the presence of the vehicle. Figure 2 shows the main components of inductive loop detectors.

Single loop detectors are capable of measuring flow and lane occupancy directly, while measurement of vehicle speed requires using dual loops or estimation using traffic flow models. Estimation of speed using traffic flow models is explained below.

Flow = speed * density…….. (1)

Where density can be approximated from lane occupancy using:

Density= occupancy * g……. (2)

and

g = k/ (vehicle length + detector length)…….(3)

Where

k is a conversion factor.

Hall and Persaud (1989) came up with different values of g for different traffic conditions.

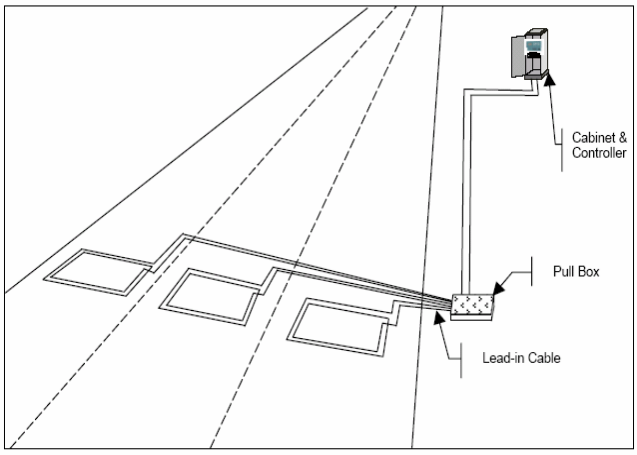


Figure 2: Inductive Loop Detectors

Speed is also estimated using dual loops and can be calculated from the formula given below. Figure 3 represents time-space diagram of the vehicle passing over two closely spaced detectors.

$$S = \frac{D}{[(t_{on})_n]_B - [(t_{on})_n]_A}$$

Where,

S is the speed of the vehicle,

D is the distance from upstream edge of detection zone A to the upstream edge of detection zone B (feet),

$[(t_{on})_n]_B$ is the instant that vehicle is detected on detector B,

$[(t_{on})_n]_B$ is the instant that vehicle is detected on detector A.

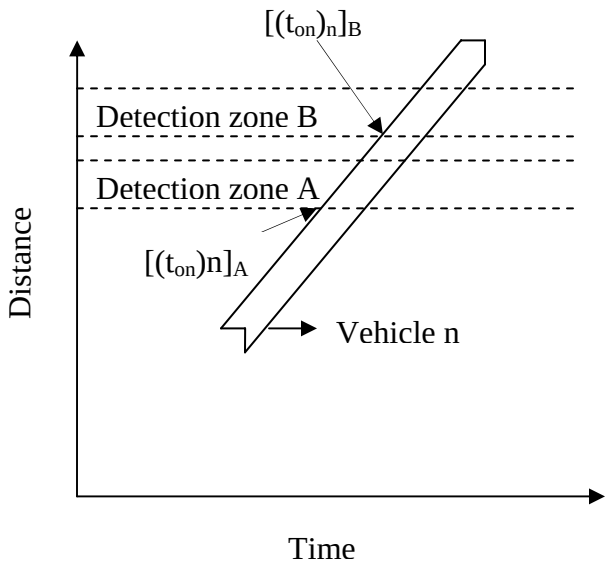


Figure 3 : Vehicle Passing over Two Closely Spaced Detectors

**2.3 Other Traffic Monitoring Devices**

Road-based traffic surveillance systems may also include other types of monitoring devices such as Closed Circuit Television (CCTV), Video Image Detection Systems (VIDS), and sensors such as Remote Traffic Microwave Sensors (RTMS). CCTV's and VIDS systems are more efficient and cost effective traffic monitoring equipment that provides real-time traffic information, but are sensitive to all weather conditions. RTMS devices are cost effective and weather resistant. They detect traffic parameters on multiple lanes and are increasingly becoming popular in the last few years. Inductive loop detectors form most commonly used traffic surveillance equipment. However, the data collected from these detectors is prone to errors due to loop malfunctions such as cross-talk (interaction of magnetic fields of the closely placed loop detectors), pulse break-up (where a single vehicle registers multiple actuations as the

sensor output flickers off and back on) , stuck sensors etc. With the wide spread of such detectors there appears to be a pressing need to monitor the data quality to ensure better reliability of subsequent applications that rely on loop detector data.

### 2.4 Data Screening Methods

Several research efforts that focused on providing algorithms to screen erroneous data were reviewed and are briefly presented in this section. In general, two basic approaches have been pursued. The first approach involves processing raw signals from the loop detectors, where the sensor on-times are used to compute the volume and occupancy, which are further checked for credibility. The second approach applies reliability checks either directly on the macroscopic parameters (volume occupancy and speed) or on the traffic relationships between these parameters, usually by establishing thresholds beyond which the data represents unrealistic traffic conditions. Examples of each approach are presented next.

Chen and May (1987) suggested a methodology for data screening in which the on-time of detector is compared with the station average to determine the inconsistency in the data. The approach was criticized being sensitive to errors such as “pulse breakups” where multiple detections of a single vehicle were registered as sensor output flickers off and back on.

Another study based on the on-time of the detector was conducted by Coifman (1999). This research study emphasized the use of speed traps to identify detector errors and assessed the performance of the speed trap based on the assumption that the on-times should be same for the vehicles at free flow conditions, allowing for hard decelerations and regardless of vehicle length. The proposed study was sensitive to congested

conditions as vehicle acceleration from low speeds will cause two on-times to differ and was not applicable for congested traffic conditions.

A later study using on-times was conducted by Coifman and Dhoorjaty (2002). This study presented several detector validation tests that use event data (individual vehicle data) to identify detector errors both at single and dual loop detectors. Detector errors were identified by a series of eight detector validation tests which used event data like head way, vehicle length, number of congested samples etc in combination to the on-time, thus making the approach applicable to all traffic flow conditions.

Examples of the studies that used the second approach include a study *by* Jacobson et al. (1990) to develop a screening algorithm based upon threshold values of occupancy, and occupancy to volume (O/V) ratios. The observations were screened with the thresholds designed to represent different malfunctioning states of the detectors and thus the observations were identified as erroneous.

Another study by Cleghorn et al (1991) suggested a data screening algorithm based upon two strategies: (i) upper bound developed for flow-occupancy data for single loop detector systems and (ii) boundaries for feasible combinations of speed, flow and occupancy data. This study indicated that erroneous observations of the traffic data could lead to deterioration of performance of incident detection algorithm.

A later study by Payne and Thompson (1998) presented various types of malfunction identification tests by imposing thresholds on occupancy, speed and volume parameters. The malfunctions were then diagnosed by inspection of aggregate sensor measurements. Data repair of faulty observations was then done by estimating actual traffic conditions and utilizing measurements from adjacent lanes.

Turochy and Smith (2000) presented a study emphasizing the development of data screening algorithm based on the combination of threshold value tests and traffic flow theory. The screening procedure devised in this research study was based on four tests. The first two tests were based on maximum volume and occupancy thresholds, while the third test was based on the maximum value of volume that could be observed for zero value of occupancy and the fourth test was based on feasibility of average vehicle length calculated as a function of speed, volume and occupancy.

Peeta and Anastassopoulos (2002) conducted a study to detect the errors due to malfunctioning detectors and predict actual data using Fourier transformation based correction heuristic. This approach was capable of detecting abnormalities and distinguishes data faults from incidents and aids the operation of online architectures of real-time route guidance and incident detection.

Ishak (2003) presented the concept of fuzzy clustering to measure the level of uncertainties associated with the three traffic parameters: speed, occupancy and volume. This research study criticized the use of average effective vehicle lengths for identifying detector data errors and devised a data screening algorithm based upon the uncertainty measure derived from membership grade and a decaying function. The uncertainty measure was then compared to a certain threshold limit to screen the observations and identify erroneous nature of single parameter.

Chen et al. (2003) developed a diagnostic algorithm to identify bad loop detectors from their speed, occupancy and volume measurements using time series of many samples. About four statistics which represent the summaries of time-series were derived

and were used to decide whether the loop is bad or good. Imputation of the missing values was done based upon the linear relationship between neighboring loops.

A study by Wall and Dailey (2003) indicated the use of consistency of vehicle counts to judge the validity of the data for an off-line analysis. The study also suggested a methodology to correct the erroneous data by identifying properly calibrated detectors which are used as reference stations to correct the data from poorly calibrated stations.

Chilkamarri and Al-Deek (2004) presented a screening algorithm to flag out bad samples using the mathematical relationship between the flow, occupancy, speed and average vehicle length and suggested a pair-wise quadratic regression model to impute the missing data in real time. The study also proposed entropy statistic to identify the detectors which are stuck. Several other studies that used the second approach for data filtering were published. See for instance (Nihan 1997-2002).

### 2.5 Summary

Most of research conducted in this area used the traffic flow relationships or imposed thresholds on the observations to devise data screening strategies. However, no effort was made to model the stochastic relationships between the traffic parameters and develop a real-time data screening algorithm. This study aims to develop a real-time data screening algorithm by considering the probabilistic relationships between the parameters which triggers the online maintenance of the detectors as well.

# 3. DATA COLLECTION

## 3.1 Introduction

This chapter describes the procedure used to collect the data for conducting the research study. The chapter also includes information about the preliminary screening techniques that were used to remove erroneous observations that result from improper recording of the data.

The data used in this study was collected from a 38-mile freeway segment of the I-4 corridor in Orlando, Florida. Figure 4 shows the map of the study section considered that extends from west of US-192 to east of Lake Mary Blvd. Data was collected using 70 inductive dual loop detector stations that are spaced at nearly .5 miles apart in both directions (east bound and west bound) on the study section considered. Each lane has two 6' x 6' loops embedded in the pavement that are connected to a 170 type controller located in a cabinet adjacent to the road side. Table 1 shows the location and description of each detector station.

Each detector station collects 30 second observations of three traffic parameters (speed, lane occupancy, and volume counts) from all six lanes. Speed and lane occupancies are expressed as average for all vehicles with in each 30 second period, while volume represents the cumulative vehicle counts within each time period (30 seconds). The information collected from each detector station is then transmitted to the Orlando Regional Traffic management center (RTMC). Figure 5 displays the configuration of a typical loop detector station in one direction of travel. The loop detector data is collected in real time via a T1 link between the Orlando RTMC and the ITS lab at the University of Central Florida. Speed, volume counts, and lane occupancies

are downloaded and compiled into an Structured Query Language (SQL) server that supports multiple publicly accessible web applications such as real-time and short-term travel time predictions between user-selected on- and off- ramps. Information about the three traffic parameters was extracted from 130 million observations that were compiled in the year 2000 and 2002.

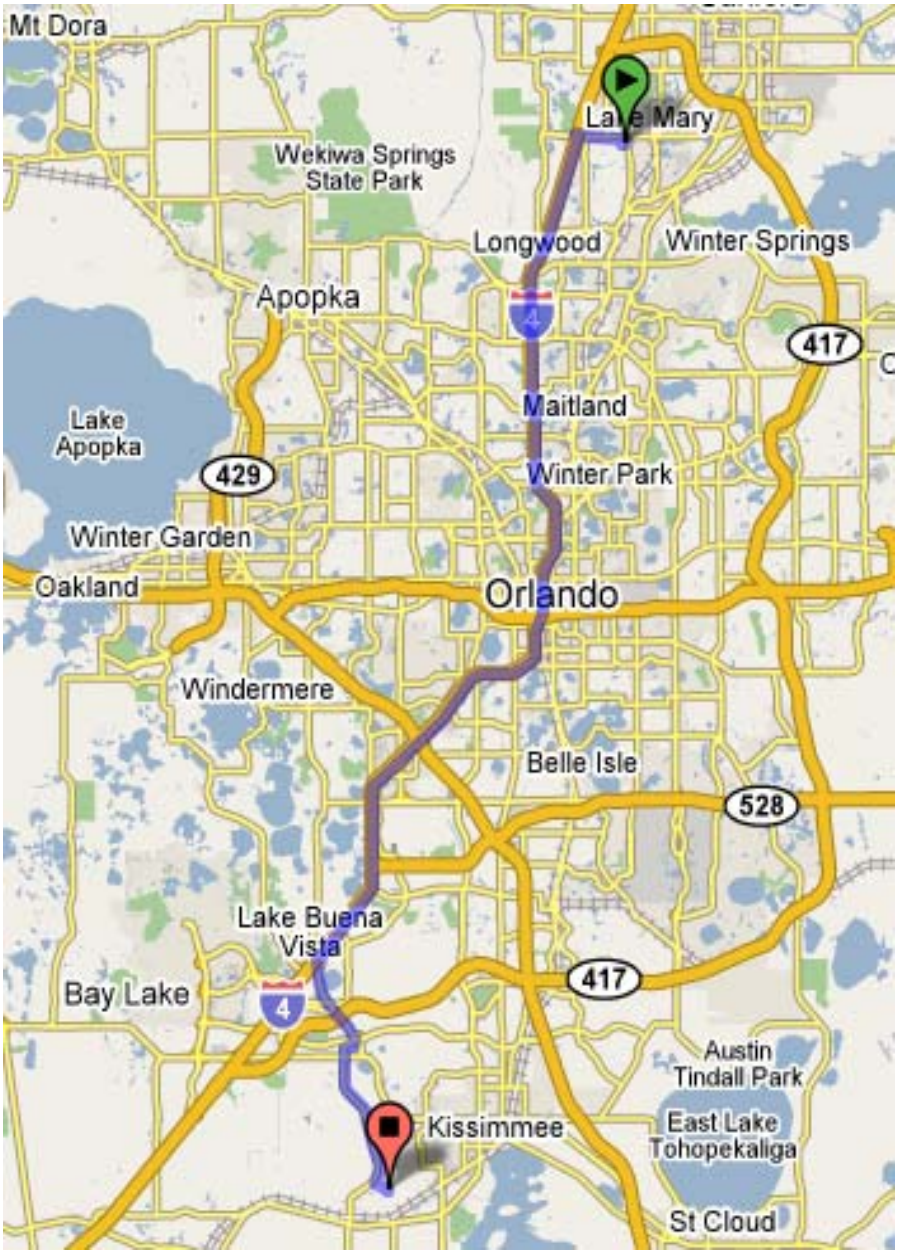

Figure 4: Map of I-4 Study Corridor in Orlando, Florida

Table 1**:** Location of Loop Detector Stations on the 38- Mile Corridor of I-4 in Orlando, Florida

| From Station | To Station | Location | Spacing (feet) |
|---|---|---|---|
| 1 | 2 | West of 192 | - |
| 2 | 3 | West of 192 | 2600 |
| 3 | 4 | US 192 | 2470 |
| 4 | 5 | West of Osceola | 3300 |
| 5 | 6 | East of Osceola | 3530 |
| 6 | 7 | SR 536 | 3330 |
| 7 | 8 | East of SR 536 | 3370 |
| 8 | 9 | West of SR 535 | 3360 |
| 9 | 10 | West of SR 535 | 3400 |
| 10 | 11 | SR 535 | 3000 |
| 11 | 12 | West of Rest Area | 3200 |
| 12 | 13 | Rest Area | 4090 |
| 13 | 14 | West of Central Florida Pkwy | 3020 |
| 14 | 15 | Central Florida Pkwy | 2980 |
| 15 | 16 | 528 EB Ramp | 2910 |
| 16 | 17 | 528 WB Ramp | 3250 |
| 17 | 18 | West of 482 | 3100 |
| 18 | 19 | West of 482 | 3450 |
| 19 | 20 | SR 482 | 2000 |
| 20 | 21 | West of 435 | 3100 |
| 21 | 22 | West of 435 | 2600 |
| 22 | 23 | SR 435 | 3000 |
| 23 | 24 | 435 WB Ramp | 2900 |
| 24 | 25 | Turnpike | 2200 |
| 25 | 26 | Turnpike WB Ramp | 2900 |
| 26 | 27 | Camera 21 | 2610 |
| 27 | 28 | West of John Young Pkwy | 2890 |
| 28 | 29 | West of John Young Pkwy | 2900 |
| 29 | 30 | John Young Pkwy | 4100 |
| 30 | 31 | East of John Young Pkwy | 2400 |
| 31 | 32 | Rio Grande | 2600 |
| 32 | 33 | Orange Blossom Trail | 2400 |
| 33 | 34 | Michigan | 2500 |
| 34 | 35 | Kaley | 2400 |

Table 1: (Continued)

| From Station | To Station | Location | Spacing (feet) |
|---|---|---|---|
| 35 | 36 | Camera 28 | 2700 |
| 36 | 37 | Camera29 | 2700 |
| 37 | 38 | Church St | 1800 |
| 38 | 40 | Robinson | 3000 |
| 39 | 41 | SR 50 | 2500 |
| 40 | 42 | Ivanhoe | 2600 |
| 41 | 43 | Princeton | 2700 |
| 42 | 44 | Winter Pk | 2600 |
| 43 | 45 | Par Ave | 2600 |
| 44 | 46 | Minnesota | 3000 |
| 45 | 47 | SR 426 | 2200 |
| 46 | 48 | Site 1393 | 2300 |
| 47 | 49 | Lee Rd | 2600 |
| 48 | 50 | East of Lee Rd | 1700 |
| 49 | 51 | Kennedy | 2800 |
| 50 | 52 | 414 EB Ramp | 3000 |
| 51 | 53 | East of SR 414 | 1800 |
| 52 | 54 | Wymore | 3300 |
| 53 | 55 | East of Wymore | 2700 |
| 54 | 56 | West of SR 436 | 2900 |
| 55 | 57 | SR 436 | 2400 |
| 56 | 58 | West of SR 434 | 3800 |
| 57 | 59 | West of SR 434 | 2900 |
| 58 | 60 | SR 434 | 3500 |
| 59 | 61 | 434 Ent Ramp | 3400 |
| 60 | 62 | 434 Ext Ramp | 1900 |
| 61 | 63 | West of EEWill | 2800 |
| 62 | 64 | East of EEWill | 2600 |
| 63 | 65 | Rest Area | 3000 |
| 64 | 66 | East of Rest Area | 2700 |
| 65 | 67 | West of Lake Mary Blvd | 2100 |
| 66 | 68 | West of Lake Mary Blvd | 2500 |
| 67 | 69 | Lake Mary | 2800 |
| 68 | 70 | Lake Mary | 2300 |
| 69 | 71 | East of Lake Mary Blvd | 3500 |

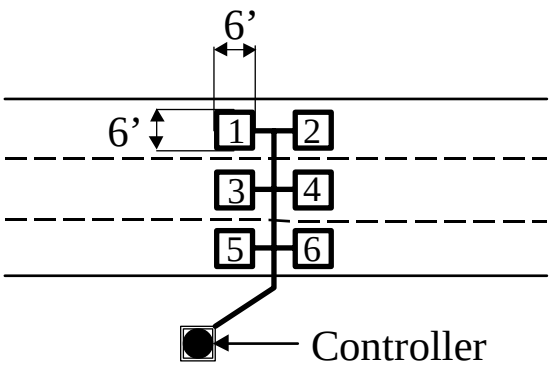


Figure 5: Typical Loop Detector Station

A sample of the data collected from the database is in the form as shown in Figure 6. The data contains the following information: station, time, and the three traffic parameters on all the six lanes.

### 3.2 Initial Data Screening

The data obtained from the loop detector could not be directly used for capturing the probabilistic relationships between the parameters. This was due to discrepancies observed in the data such as negative value or error that resulted from the break down of the detector station or failure of communication infrastructure between detector station and TMC. Hence the data had to be filtered to remove these erroneous observations. Each dual loop detector records two values of occupancy, and volume count that were averaged, while speed is directly calculated using the dual loops. Preliminary filtering techniques that were used to remove invalid observations of the three traffic parameters are listed as follows:

- Observations with zero or negative values of the parameters were discarded.

- The errors in the data resulting from the failure of the loop detectors or mis-functioning of the communication between detector and TMC are represented by a –9XX value,-XX value or zero and are filtered out.

### 3.3 Summary

Information in terms of three macroscopic traffic parameters (occupancy, speed and volume was collected from dual loop detection system on the study section. The data was then processed to filter out the preliminary errors (such as negative errors or mis-communication errors) associated with the data. The data can now be used to examine the probabilistic relationships between the parameters.

| station | time | els | ecs | ers | wls | wcs | wrs | elv | ecv | erv | wlv | wcv | wrv | elo | eco | ero | wlo | wco | wro |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 06:30.0 | 0 | 51 | 56 | 0 | 63 | 0 | 0 | 6 | 10 | 0 | 2 | 0 | 0 | 4 | 7 | 0 | 1 | 0 |
| 2 | 07:00.0 | 57 | 57 | 66 | 66 | 54 | 48 | 4 | 4 | 4 | 1 | 4 | 2 | 4 | 2 | 2 | 0 | 3 | 1 |
| 2 | 07:30.0 | 9 | 56 | 0 | 60 | 70 | 71 | 2 | 7 | 0 | 4 | 2 | 1 | 0 | 2 | 0 | 1 | 1 | 0 |
| 2 | 08:00.0 | 59 | 51 | 59 | 57 | 61 | 61 | 1 | 2 | 1 | 2 | 4 | 2 | 0 | 0 | 0 | 0 | 1 | 0 |
| 2 | 08:30.0 | 58 | 60 | 82 | 77 | 70 | 100 | 5 | 8 | 3 | 2 | 5 | 4 | 3 | 5 | 2 | 1 | 4 | 2 |
| 2 | 09:00.0 | 57 | 59 | 61 | 61 | 60 | 63 | 3 | 4 | 4 | 1 | 2 | 1 | 2 | 2 | 2 | 0 | 1 | 0 |
| 2 | 09:30.0 | 56 | 59 | 63 | 0 | 51 | 59 | 8 | 5 | 1 | 0 | 3 | 3 | 3 | 2 | 0 | 0 | 2 | 1 |

[1]

Figure 6: Sample of SQL Compiled Data for January 2000

[1] els, ecs, ers - speed in the east bound direction on left, center and right lanes. wls, wcs and wrs-speed in west bound direction on left, center and right lanes. elo, eco, ero- occupancy in east bound direction on left, center and right lanes. wlo, wco, wro- occupancy in west bound direction on left, center and right lanes. elv, ecv, erv- volume in east bound on left, center and right lanes. wlv, wcv, wrv- volume in west bound on left, center and right lanes.

# 4. METHODOLOGY

## 4.1 Introduction

This research study proposed probabilistic approaches for real-time freeway traffic data screening. The proposed approaches differ from the deterministic approach in that they do not explicitly confirm the validity of an observation but attempt to quantify the likelihood that such observation is valid. Probabilistic relationships between the three traffic parameters (volume, speed, and lane occupancy) were developed to capture the least likely temporal changes in traffic states as well as inconsistencies in traffic conditions expressed by each traffic parameter. The proposed methodology thus primarily investigates the stochastic variation in traffic conditions over time and the probabilistic relationships between the three traffic parameters in order to capture certain characteristics that can be used for data screening purposes.

The methodology is derived from two complimentary approaches. The first approach considers the stochastic evolution of traffic conditions over time measured by each of the three parameters independently. The second approach attempts to capture the inherent stochastic variation of traffic conditions measured by each combination of the three traffic parameters. In both approaches models for the conditional probabilities are developed from a vast amount of detector data describing all possible variations of traffic conditions on the study segment considered. Both approaches form the basis for the data screening algorithm and are explained in detail next.

## 4.2 Approach One: Examining Temporal Variations of Traffic Parameters

This approach focuses on capturing possible abrupt changes in traffic conditions

that may occur between two successive observations taken over a time span of 30 seconds. These temporal variations, though abrupt, are unlikely to be extreme within the time span considered. For instance, the variation in the value of speed from 90 mph to 0 mph over a time span of 30 seconds is likely to be unrealistic. Thus this approach checks for unrealistic temporal variations that could be used to judge the validity of an observation. In simple terms, an observation is considered valid if the temporal variation from its preceding observation is not unrealistic. The feasible range of temporal variations could be derived by comparing the probabilities of the temporal variations with user-specified thresholds.

The temporal changes observed between the parameters over time are stochastic due to random variations in the traffic conditions. The stochastic variation of the parameters is captured using the conditional probability concept as mentioned earlier. Conditional probability is defined as the probability of an event occurring given that some event has occurred (see Myers, 2002).

The methodology used to examine stochastic variations of each traffic parameter over time requires estimation of family of cumulative PDFs. Let $X_t$ and $X_{t+1,}$ represent any of the three parameters (speed, occupancy and volume) observed at time t and t+1. The difference $(X_t - X_{t+1})$, refers to the drop or increase in $X$ over a duration of 30 seconds. Several possible combinations of two successive observations were used to model all possible temporal variations for each variable. The probability of observing the difference $(X_t - X_{t+1})$, given $(X_t = x)$ is estimated using the following discrete conditional probability density function:

$$P\left\{X_t - X_{t+1} = \delta \mid x\right\} = \frac{N\left\{X_t - X_{t+1} = \delta \mid x\right\}}{N\left\{x\right\}} \quad (4)$$

Where,

$\delta$ is the realization of the random variable $X_t - X_{t+1}$,

$P\left\{X_t - X_{t+1} = \delta \mid x\right\}$ = the probability of observing a difference of $\delta$ given $X_t = x$,

$N\left\{X_t - X_{t+1} = \delta \mid x\right\}$ = the number of observations with the difference of $\delta$ given, $X_t = x$,

$N\left\{x\right\}$ = total number of observations with $X_t = x$.

The difference between the variables ($X_t - X_{t+1}$) of two successive observations may be positive or negative, indicating either a drop or an increase in the value of $X$ over a time interval of 30 seconds. The probability distribution functions for drops or increases in X exhibit different characteristics. Hence, they are studied separately. The probability distribution function for a *drop* in ($X$) can be found from the cumulative sum of discrete probability mass function as follows:

$$P\left\{X_t - X_{t+1} \le \delta \mid x\right\} = \sum_{\forall j \in [0,\delta]} P\left\{X_t - X_{t+1} = j \mid x\right\} \quad \forall x \in \left\{0, X^{\max}\right\}, \quad \delta \ge 0 \quad (5)$$

Where,

$P\left\{X_t - X_{t+1} \le \delta \mid x\right\}$ = the cumulative probability of observing a drop in $X$, given $X_t = x$,

$X^{\max}$ is the maximum feasible value for variable $X$.

The probability distribution function for *increase* in ($X$) can be similarly found from the cumulative sum of discrete probability mass function as follows:

$$P\left\{X_{t+1} - X_t \le \delta \mid x\right\} = \sum_{\forall j \in [0,\delta]} P\left\{X_{t+1} - X_t = j \mid x\right\} \quad \forall x \in \left\{0, X^{\max}\right\}, \quad \delta \ge 0 \quad (6)$$

Where,

$P\{X_{t+1} - X_t \le \delta \mid x\}$ = the cumulative probability of observing an increase in $X$, given $X_t = x$ and $\delta \ge 0$.

### 4.3 Approach Two: Examining Probabilistic Traffic Flow Relationships

This approach aims at checking for inconsistencies in the traffic conditions measured by each of the traffic parameters. Volume count, lane occupancy and speed in any observation are inter-related and should reflect similar traffic conditions. The relationship between the three traffic parameters can be used as a measure to validate an observation. For example, an observation representing a combination of high speed and high occupancy is unlikely under stable flow conditions. Such combinations are inconsistent, and possess less probability. Thus examining the probabilistic relationship between the parameters serves as a source for detecting the inconsistencies in the traffic conditions.

The relationship between the parameters is however probabilistic due to the random changes in the traffic conditions. These relationships are examined using conditional probability concept as mentioned earlier. Estimation of PDFs that represent the probabilistic relationship between the three traffic parameters is described next. Let variables X and Y represent two of the three traffic parameters (speed, occupancy and volume). The probabilistic relationship between X and Y may be approximated by a probability mass function of the form:

$$P\{X = x_i \mid Y = y_j\} = \frac{N\{X = x_i \mid Y = y_j\}}{N\{Y = y_j\}} \qquad \forall\ i \in [0, N],\ j \in [0, M] \qquad (6)$$

Where

$N$ is the number of realizations of $X$ .

$M$ is the number of realizations of $Y$ .

$P\left\{X = x_i \mid Y = y_j\right\}$ = the conditional probability of observing $X = x_i$ given $Y = y_j$ .

$N\left\{X = x_i \mid Y = y_j\right\}$ = the number of observations of $X = x_i$ given $Y = y_j$ .

$N\left\{Y = y_j\right\}$ = total number of observations with $Y = y_j$ .

The discrete probability function is used to calculate the cumulative distribution function as follows:

$$P(X \le x_k \mid Y = y_j) = \sum_{\forall i \in [0,k]} P(X = x_i \mid Y = y_j) \qquad \forall j \in [0, M],\ K \in [0, N]. \tag{7}$$

Where

$P(X \le x_k \mid Y = y_j)$ is the probability of observing $X \le x_k$ given $Y = y_j$ .

### 4.4 Summary

Probability distribution functions that represent the temporal variations of each parameter, and the probabilistic traffic flow relationships were estimated. These PDFs functions are to be modeled to capture the nature of these relationships. This raises the issue of choosing appropriate modeling tools, which is addressed in the next section.

# 5. MODELING PROBABILITY DISTRIBUTION FUNCTIONS

## 5.1 Introduction

The probability distribution functions derived from the two approaches reflect the random behavior of the traffic conditions, and are non-linear in nature. Hence a non-linear function approximation seems quite appropriate to model the data. This can be best accomplished using Artificial Neural Networks (ANN).This chapter presents an introduction to Multi-Layer Perceptron (MLP), an Artificial Neural Network (ANN) tool used for function approximation. This chapter also explains the procedure to train NN models for approximating the probability distribution functions. Finally the performance evaluation of network models built is presented.

## 5.2 Multi-Layer Perceptron (MLP)

The MLP is a general static ANN that has been used extensively for nonlinear function approximation. It consists of four layers- input layer, two hidden layers and an output layer. Input layer where the data is fed; the hidden layers which extract the features from the input patterns; and the output layer which gives the responses to the input fed into the network. MLP is trained using back propagation algorithm, which minimizes the sum of squared errors between the desired and actual output. Figure 7 shows an example of network topology used for the study. The number of neurons in the first hidden layer is double the number of neurons in the second one, as a general practice in NNs topology.

## 5.3 Modeling PDFs for Approach One

The process of approximating the discrete probability distribution functions developed from the two approaches is done by training MLP networks with the data.

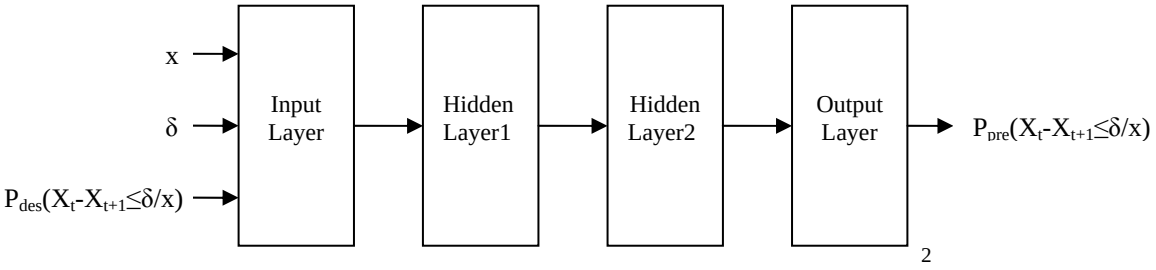


Figure 7: An Example of MLP Network Topology

The probability distribution functions estimated for each traffic parameter (for both drop and increase conditions) were approximated separately using different MLP networks as they possessed different characteristics and probability distributions. The probability distribution function for drop in X was expressed as follows:

$$P\{X_t - X_{t+1} \leq \delta \mid x\} = \sum_{\forall j \in [0,\delta]} P\{X_t - X_{t+1} = j \mid x\} \quad \forall x \in \{0, X^{\max}\}, \quad \delta \geq 0$$

The input data for modeling the PDF representing a drop in X was in the following form:

Type 1 Input: $\{X, \delta, P\{X_t - X_{t+1} \leq \delta \mid x\}\}$

Where,

$X_t$ and $X_{t+1}$, represent any of the three parameters (speed, occupancy and volume) observed at time t and t+1,

$\delta$ is the realization of the random variable $X_t - X_{t+1}$,

$P\{X_t - X_{t+1} \leq \delta \mid x\}$ = the cumulative probability of observing a drop in $X$, given $X_t = x$.

The probability distribution function for increase in X was expressed as follows:

$$P\{X_{t+1} - X_t \leq \delta \mid x\} = \sum_{\forall j \in [0,\delta]} P\{X_{t+1} - X_t = j \mid x\} \quad \forall x \in \{0, X^{\max}\}, \quad \delta \geq 0$$

[2] $P_{des}$- probability desired $P_{pre}$- probability predicted

The input data for modeling the PDF representing an increase in X was in the following form:

Type 2 Input: $\{ X, \delta \ , \ P\{ X_{t+1} - X_t \le \delta \mid x \} \}$

Where,

$\delta$ is the realization of the random variable $X_{t+1} - X_t$,

$P\{ X_{t+1} - X_t \le \delta \mid x \}$ = the cumulative probability of observing an increase in $X$, given $X_t = x$ and $\delta \ge 0$.

Separate data sets for modeling the stochastic variations of each traffic parameter were extracted from a large data set complied in the year 2000. These data sets were used to train the MLP networks. Speed and occupancy parameters varying from a range of (0-100 mph) and (0-100 %) were considered to account for the most likely traffic conditions. The maximum value for the volume count was taken to be 20 in compliance with the maximum capacity of 2400 vphpl. Cumulative probabilities representing the stochastic variations for each traffic parameter were calculated using the PDFs. Figure 8 shows a sample of the probability distributions for the stochastic temporal variation of three traffic parameters. These PDFs were now approximated using ANN by inputting the data in the format mentioned earlier as explained next.

Multi Layer Feed-forward networks were trained with the input data using back propagation algorithm to capture the stochastic variation of the parameters over time. The input data set consists of two independent variables and a dependent variable. The number of neurons in the input layer depends on the number of independent variables considered. Hence the number of neurons in the input layer was fixed to two. The output layer contains a single neuron that represents the dependant variable. The number of

neurons in the hidden layers was arbitrarily chosen depending upon the size of the training data. Training process progresses with the aim of reducing the mean square error on the training data and was terminated on the basis of any of the two criteria (i) training error reaching minimum (i.e. MSE is equal to .01) or (ii) training epochs reaching a maximum of 1000.

Occupancy and speed parameters ranging from (zero to hundred) were split into four uniform intervals of twenty five each, and their corresponding stochastic variations were approximated using separate networks to improve approximation efficiency.

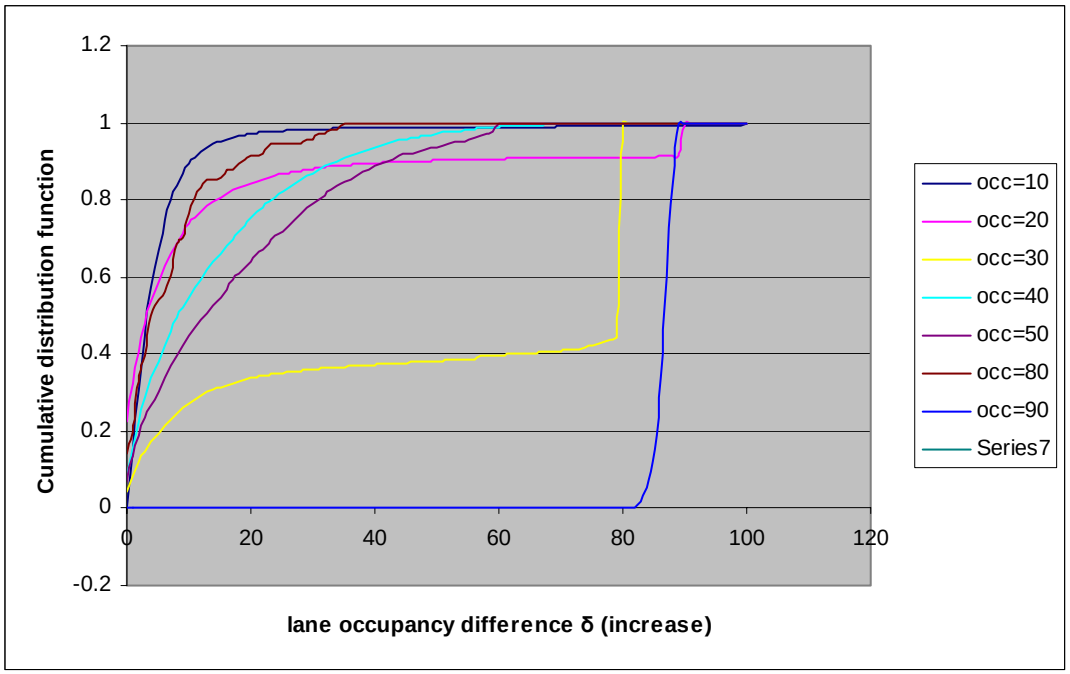


Figure 8: Probability Distribution Functions for Occupancy Parameter

Stochastic variations of volume parameter were modeled directly as the approximation performance achieved was observed to be high. Eight networks (four networks for stochastic variations representing drop and four networks for stochastic variations representing increase) were built to model stochastic variations of either

occupancy or speed parameter. Two MLP networks were built to model the stochastic variation of volume parameter.

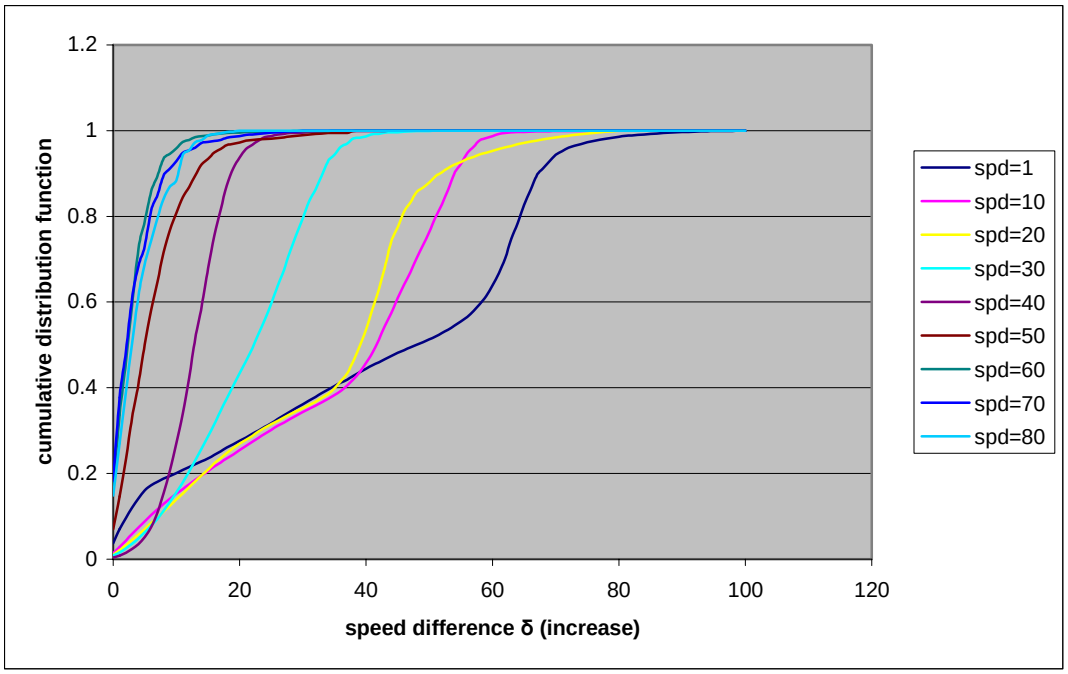


Figure 9: Probability Distribution Functions for Speed Parameter

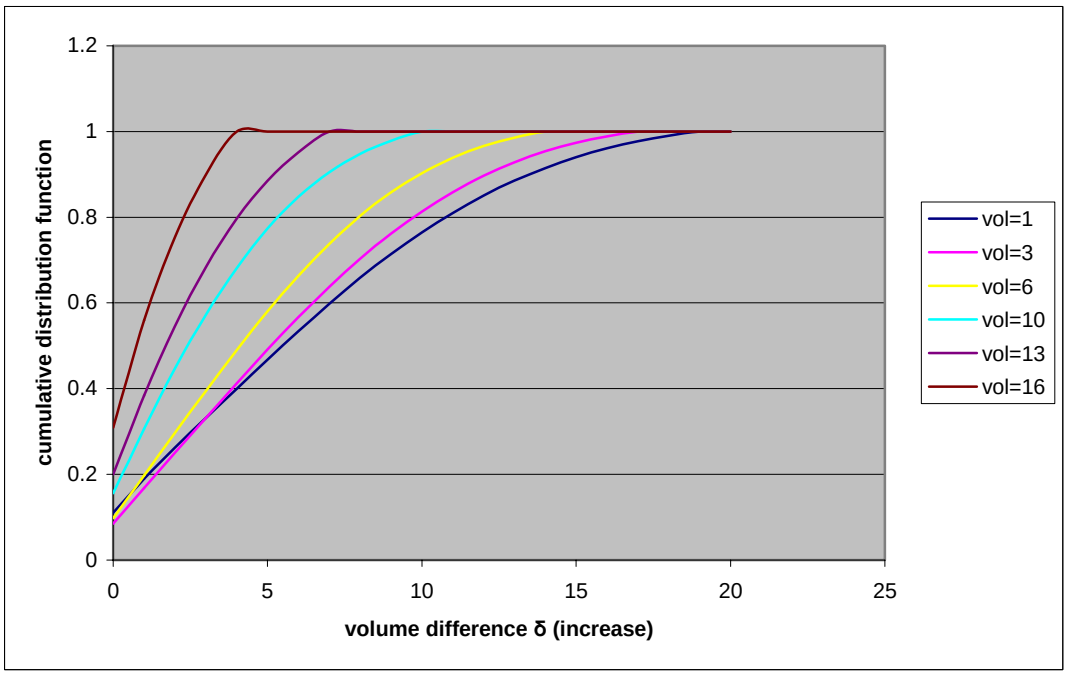


Figure 10: Probability Distribution Functions for Volume Parameter

A total of eighteen MLP networks ((2*8) for occupancy and speed parameters + 2 for volume parameter) were built to approximate the stochastic variation of the three traffic parameters.

### 5.4 Modeling PDFs for Approach Two

Probability distribution functions representing relationship between the parameters were approximated separately using MLP networks. The probability distribution function for relationship between any two parameters was expressed as follows:

$$P(X \le x_k \mid Y = y_j) = \sum_{\forall i \in [0,k]} P(X = x_i \mid Y = y_j) \qquad \forall j \in [0, M],\ K \in [0, N].$$

The input data for modeling the PDF above was in the following form:

Type 1 input: $\{X, Y, P(X \le x_k \mid Y = y_j)\}$

Where,

X and Y represent two of the three traffic parameters,

$P(X \le x_k \mid Y = y_j)$ is the probability of observing $X \le x_k$ given $Y = y_j$,

$N$ is the number of realizations of $X$,

$M$ is the number of realizations of $Y$.

Speed and occupancy parameters were divided into bins of 5 interval size. Separate data sets were extracted for speed conditioned on occupancy, occupancy conditioned on speed, volume conditioned on speed and volume conditioned on occupancy and the corresponding probabilities were calculated from the probability distribution functions. For the cases, speed conditioned on volume and occupancy conditioned on volume, the data was divided into stable and unstable flows. This was due

to fact that each value of volume corresponded to two values of speed or occupancy, one in the stable flow and other in unstable flow conditions. These observations possessed different probability distributions and thus they were to be modeled separately. Critical speed which separates stable flow and unstable flow conditions was calculated from weighted average method and was found to be varying between 35-40 mph. The data set for speed conditioned on volume was divided using this critical speed and the cumulative probabilities were calculated separately. Similarly critical occupancy (15-20%) which demarcates the stable flow from the unstable flow was calculated using weighted average method and their cumulative probabilities were calculated separately. The Figures from 11 to 18 show the probability distributions of several combinations of the parameters.

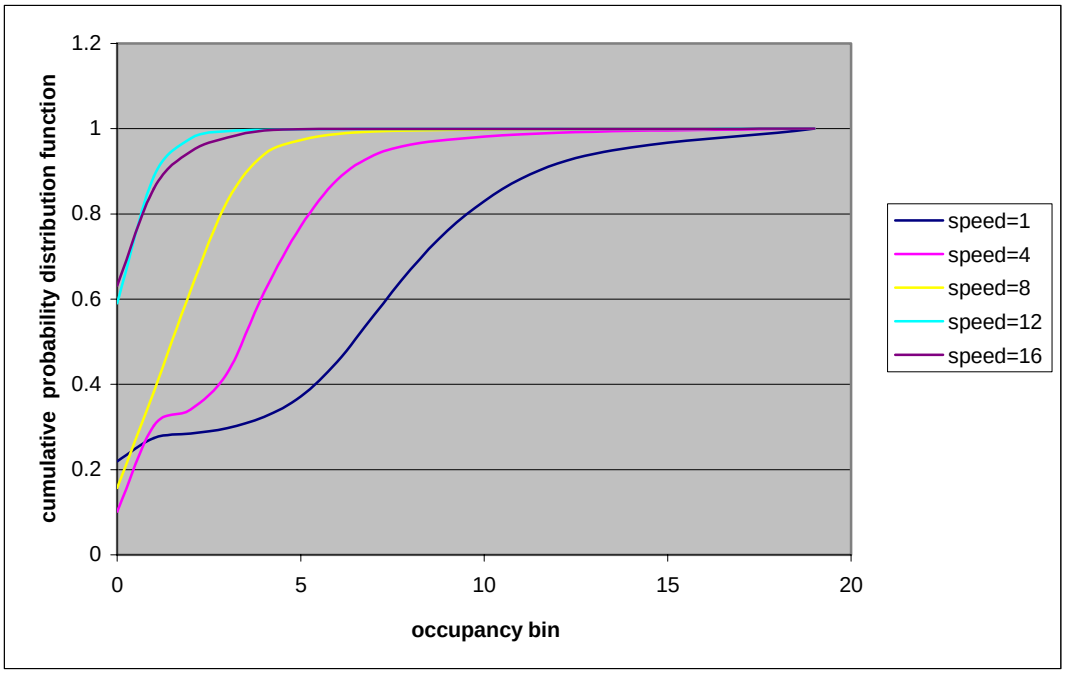


Figure 11: Probability Distribution Functions for Occupancy Conditioned on Speed

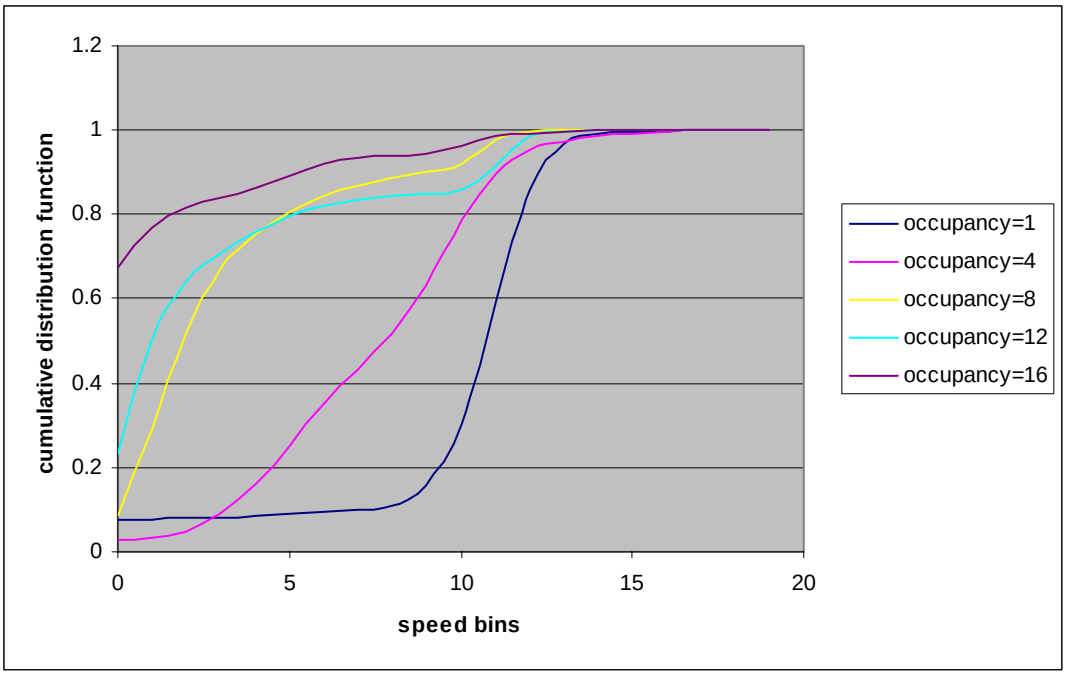


Figure 12: Probability Distribution Functions for Speed Conditioned on Occupancy

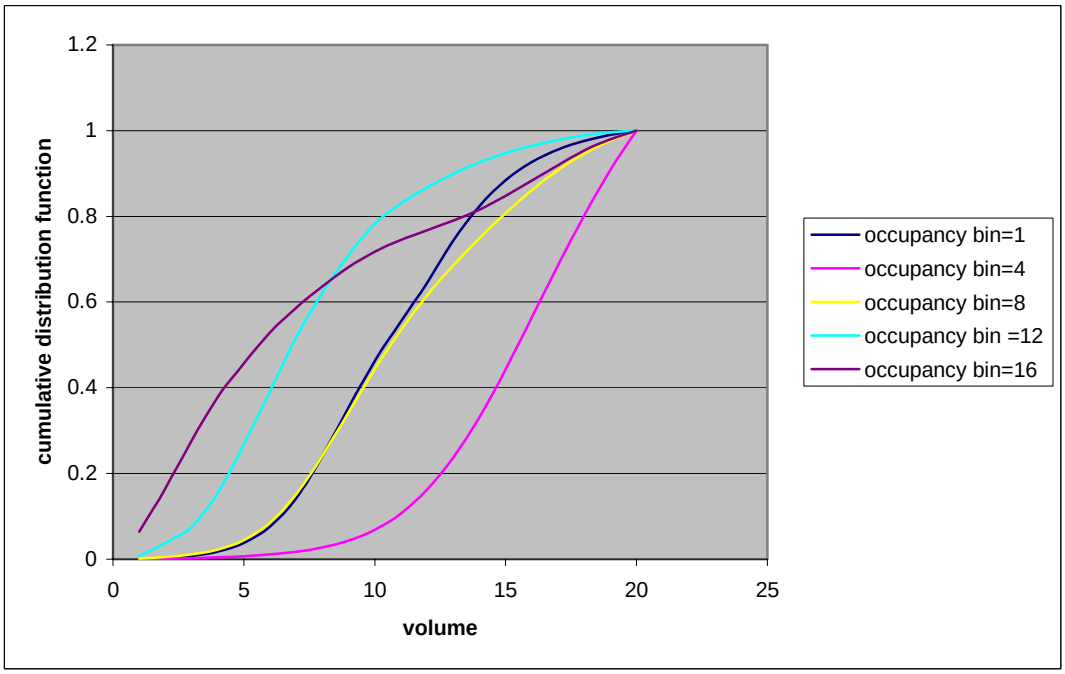


Figure 13: Probability Distribution Functions for Volume Conditioned on Occupancy

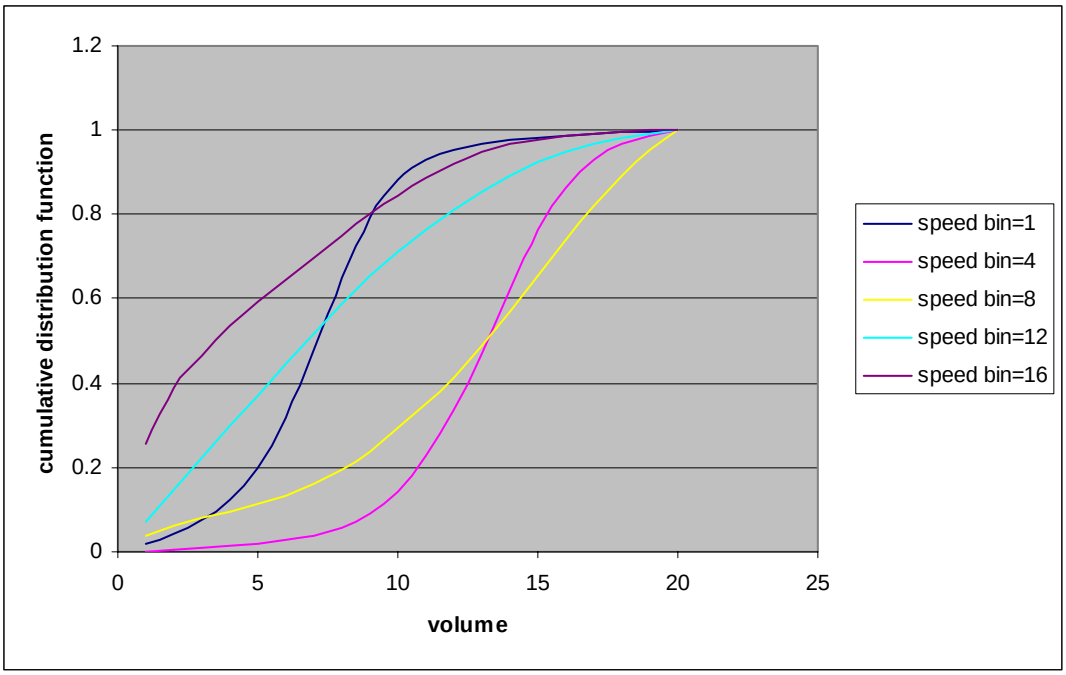


Figure 14: Probability Distribution Functions for Volume Conditioned on Speed

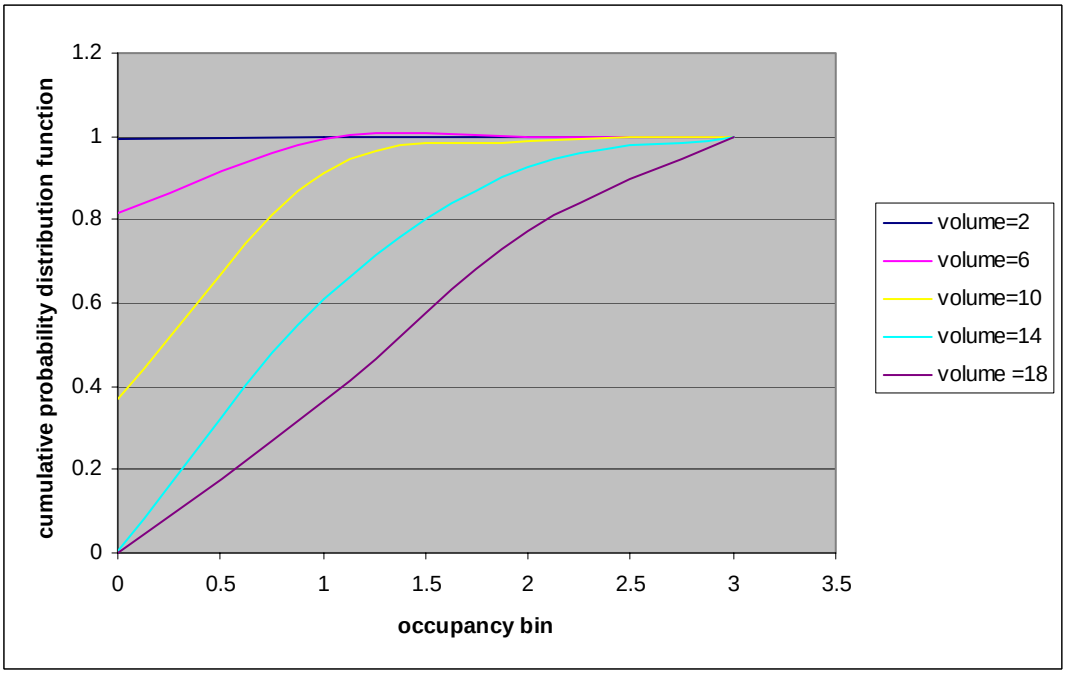


Figure 15: Probability Distribution Functions for Occupancy Conditioned on Volume (Stable flow)

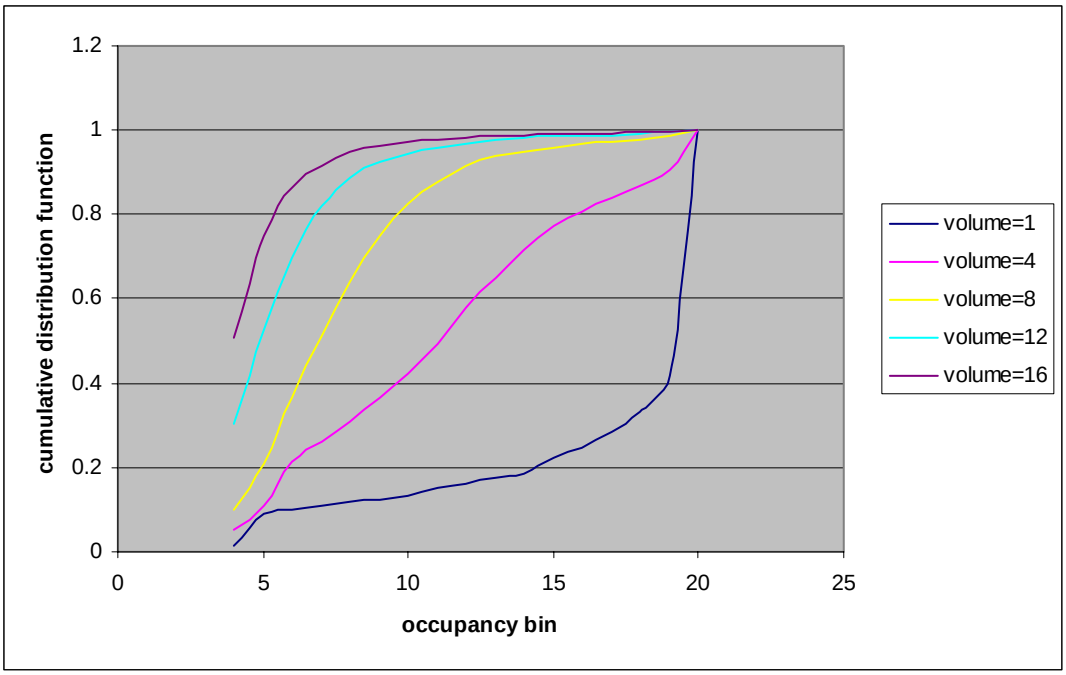


Figure 16: Probability Distribution Functions for Occupancy Conditioned on Volume (Unstable flow)

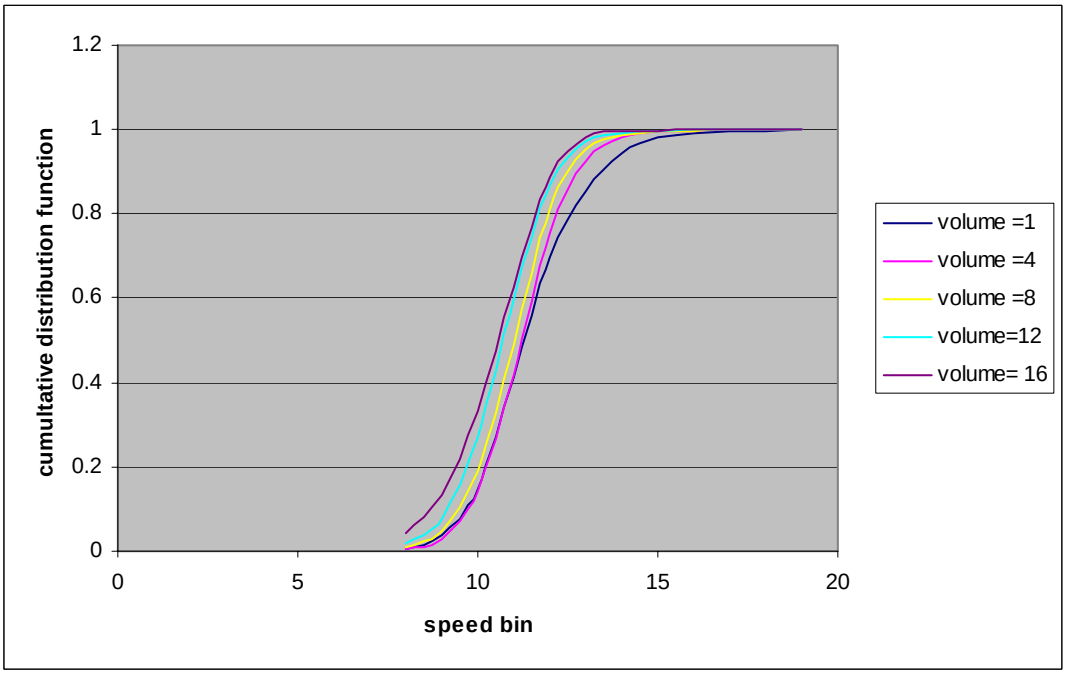


Figure 17: Probability Distribution Functions for Speed Conditioned on Volume (Stable flow)

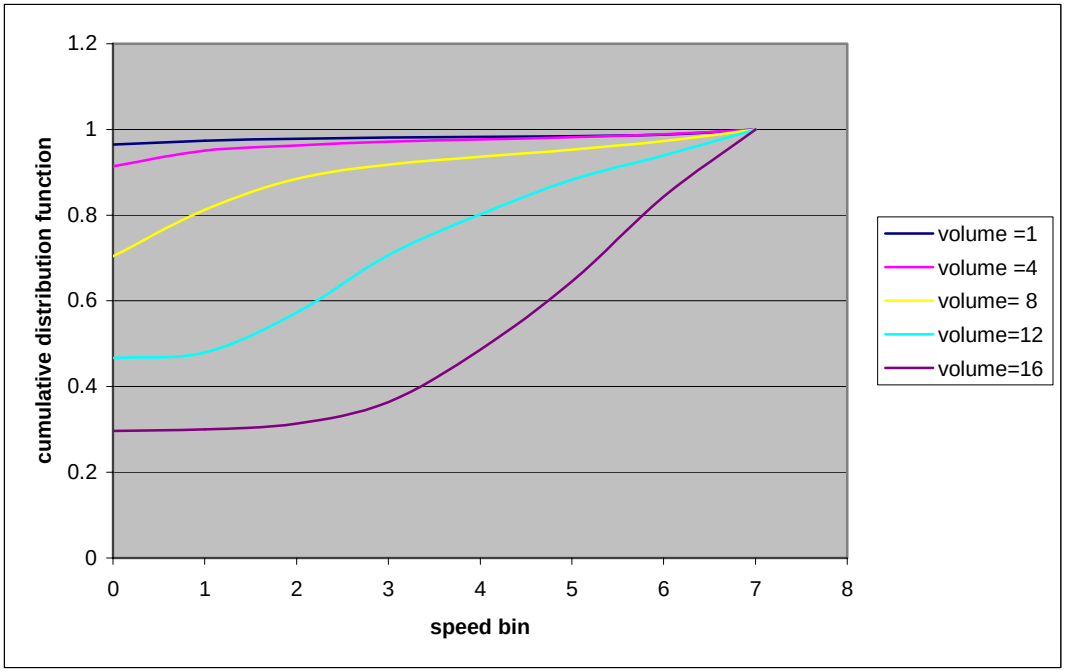


Figure 18: Probability Distribution Functions for Speed Conditioned on Volume (Unstable flow)

Multi-layer feed forward networks were trained with the input data in order to approximate the probabilistic traffic flow relationships. The data sets representing combinations of speed conditioned on occupancy, volume condition on occupancy, occupancy conditioned on speed and volume conditioned on speed were divided into four uniform intervals and were approximated using different networks. A total of 16 MLP networks (4 for each condition, therefore 16 for all the four combinations) were built to model the probabilistic relationships between the above combinations of parameters. The data sets representing the combinations of speed conditioned on volume and occupancy conditioned on volume was divided into stable flow and unstable flow conditions and were approximated using 4 MLP networks (2 for each combination, therefore 4 for two combinations.) A total of 20 networks were built to approximate the PDFs representing probabilistic traffic flow relationships.

### 5.5 Performance Measures

Trained networks are evaluated using a set of performance measures. Desired probabilities are compared with the predicted probability values generated from the network to calculate three measures of performance: R-square, and Root Mean Square Error (RMSE) and Average Absolute Relative Error (AARE). Each measure of performance is defined as below:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (p(i) - O(i))^2}{N}}$$

And

$$\text{AARE} = \frac{\sum_{i=1}^{n} \left| \frac{P(i) - O(i)}{O(i)} \right|}{N}$$

Where

$P(i)$ = predicted value of the parameter for observation i.

$O(i)$ = actual observed value of the parameter for observation i.

N = number of observations.

#### 5.5.1 Performance Evaluation of the ANN Models Developed for Approach One

The performance measures suggested above were used to evaluate the training efficiency of 18 ANN models built for approach one. Table 2 shows the performance measures for all the networks. From the Table 2 it was observed that RMSE ranges from .072 to .252, AARE varies from .0258 to .477, and R-square varies from .651 to .962. A possible explanation to the variation observed in the performance measures would be variation in the frequencies of observations leading to wide differences in the probability

distributions within the interval considered. Failure to approximate such cases accurately also leads to high approximation errors. For instance, occupancy range of 75-100 implies extreme congested conditions. A wide variation in the probability distributions of different observations could be seen due to unstable flow conditions, thus affecting the approximation efficiency. The variation in the performance measures for the occupancy increase case can be attributed to difference in the probability distributions among the observations and approximation errors.

### 5.5.2 Performance Evaluation of the ANN Models Developed for Approach Two

Performance of the 20 networks built to model the probabilistic traffic flow relationships was evaluated and tabulated in Table 3. A difference in the performance measures could be observed from Table 3 (i.e. RMSE varies from .005-.092, AARE varies from .021-.361 and R-square varies from .835-.992.). This variation in the performance measures could be attributed to variation in the probability distributions due to insufficient data and approximation errors. The performance measures calculated for evaluating the network models indicated reasonable approximation of the PDFs developed for both the approaches.

## 5.6 Summary

Probability distribution functions for first and second approach were approximated using 38 Multi-layer Feed-Forward networks. Performance evaluation of the network models conducted showed reasonable approximation of the PDFs. The network models built have the capability to predict the probabilities of real-time data which form basis for devising a data screening algorithm as explained next.

Table 2: Performance Measures of the ANN Models for Approach One

| Network No | No. of observations | Parameter | Type | Interval | RMSE | AARE | R-square | Observations within ± |
|---|---|---|---|---|---|---|---|---|
| 1 | 299 | Occupancy | Drop | 0-25 | 0.076 | 0.215 | 0.946 | 90.2 |
| 2 | 800 | Occupancy | Drop | 25-50 | 0.129 | 0.393 | 0.856 | 75.8 |
| 3 | 481 | Occupancy | Drop | 50-75 | 0.078 | 0.477 | 0.945 | 89.6 |
| 4 | 966 | Occupancy | Drop | 75-100 | 0.252 | 0.373 | 0.651 | 61.2 |
| 5 | 2162 | Occupancy | Increase | 0-25 | 0.102 | 0.114 | 0.736 | 87.6 |
| 6 | 1321 | Occupancy | Increase | 25-50 | 0.123 | 0.18 | 0.765 | 83.6 |
| 7 | 286 | Occupancy | Increase | 50-75 | 0.131 | 0.21 | 0.812 | 80.9 |
| 8 | 165 | Occupancy | Increase | 75-100 | 0.078 | 0.25 | 0.962 | 91.8 |
| 9 | 312 | Speed | Drop | 0-25 | 0.092 | 0.361 | 0.921 | 81.2 |
| 10 | 934 | Speed | Drop | 25-50 | 0.088 | 0.331 | 0.917 | 86.5 |
| 11 | 1570 | Speed | Drop | 50-75 | 0.077 | 0.124 | 0.88 | 92.2 |
| 12 | 1177 | Speed | Drop | 75-100 | 0.078 | 0.23 | 0.938 | 84.9 |
| 13 | 2039 | Speed | Increase | 0-25 | 0.092 | 0.416 | 0.946 | 83.5 |
| 14 | 1600 | Speed | Increase | 25-50 | 0.083 | 0.279 | 0.948 | 88.8 |
| 15 | 1003 | Speed | Increase | 50-75 | 0.072 | 0.113 | 0.879 | 91.1 |
| 16 | 252 | Speed | Increase | 75-100 | 0.075 | 0.252 | 0.921 | 88.4 |
| 17 | 420 | Volume | Drop | 0-20 | 0.076 | 0.0258 | 0.94 | 88.59 |
| 18 | 231 | Volume | Increase | 0-20 | 0.091 | 0.104 | 0.935 | 88.59 |

Table 3: Performance Measures of the ANN Models for Approach Two

| Network no | No of observations | Type | Interval | RMSE | AARE | R-squar | Observations within ± .01 |
|---|---|---|---|---|---|---|---|
| 1 | 100 | O/S | 0-5 | 0.092 | 0.361 | 0.921 | 90.1 |
| 2 | 100 | O/S | 5-10 | 0.088 | 0.331 | 0.917 | 95.2 |
| 3 | 100 | O/S | 10-15 | 0.077 | 0.124 | 0.88 | 95.2 |
| 4 | 100 | O/S | 15-20 | 0.078 | 0.23 | 0.938 | 96.5 |
| 5 | 81 | O/V (stable flow) | 1-3 | 0.065 | 0.193 | 0.976 | 90.1 |
| 6 | 339 | O/V (unstable flow) | 4-20 | 0.005 | 0.147 | 0.935 | 91.4 |
| 7 | 100 | S/O | 0-5 | 0.076 | 0.285 | 0.964 | 89.1 |
| 8 | 100 | S/O | 5-10 | 0.07 | 0.259 | 0.967 | 91.2 |
| 9 | 100 | S/O | 10-15 | 0.065 | 0.0907 | 0.907 | 91.2 |
| 10 | 100 | S/O | 15-20 | 0.064 | 0.021 | 0.835 | 99.1 |
| 11 | 160 | S/V (unstable flow) | 1-7 | 0.071 | 0.071 | 0.939 | 95.1 |
| 12 | 240 | S/V ( stable flow) | 8-20 | 0.066 | 0.285 | 0.992 | 94.5 |
| 13 | 100 | V/O | 0-5 | 0.079 | 0.461 | 0.969 | 88.3 |
| 14 | 100 | V/O | 5-10 | 0.074 | 0.581 | 0.982 | 85.1 |
| 15 | 100 | V/O | 10-15 | 0.075 | 0.345 | 0.967 | 86.8 |
| 16 | 100 | V/O | 15-20 | 0.077 | 0.147 | 0.919 | 88.8 |
| 17 | 100 | V/S | 0-5 | 0.072 | 0.513 | 0.973 | 92.8 |
| 18 | 100 | V/S | 5-10 | 0.078 | 0.502 | 0.952 | 90.1 |
| 19 | 100 | V/S | 10-15 | 0.061 | 0.208 | 0.977 | 90.9 |
| 20 | 100 | V/S | 15-20 | 0.066 | 0.091 | 0.954 | 94 |

# 6. DATA SCREENING ALGORITHM

## 6.1 Introduction

This chapter deals with application of the neural network models for screening of a real-time data set. The process of devising a real-time screening algorithm is carried out in three stages. In the first stage, ANN models developed are used to predict the probabilities for real-time data. In the second stage, the probabilities predicted from the network models are compared with user specific threshold to identify erroneous observations. The third stage deals with further analysis conducted to identify the erroneous parameters in an observation.

## 6.2 Stage One: Prediction of Probabilities for Real-time Data

The neural network models built were used to predict the probabilities associated with the data presented in real-time format. A twenty-four hour detector data compiled from seventy detector stations in the year 2002 was considered. A continuous stream of observations was extracted and was inputted into the network models. The probabilities associated with 52000 continuous observations were predicted from the network models and were further used for screening analysis. Figure 19 shows a snapshot of nine probabilities derived to screen each observation.

## 6.3 Stage Two: Data Screening Algorithm

The probabilities obtained from the MLP networks were used for deriving data screening strategy to filter the observations. A threshold of 95% was considered to demonstrate the implementation of data screening algorithm. The threshold specified is the probability with which the observations could be judged as valid.

| | | | | Probabilisic traffic flow relationships | | | | | | Temporal variation of the parameters | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| obs | o | s | v | $P(O \le o_k / S = s_j)$ | $P(O \le o_k / V = v_j)$ | $P(S \le s_j / O = o_j)$ | $P(S \le s_j / V = v_j)$ | $P(V \le v_i / O = o_j)$ | $P(V \le v_i / S = s_j)$ | $P\{O_t - O_{t+1} \le \delta / o\} / P\{Ot+1 - Ot \le \delta / o\}$ | $P\{S_t - S_{t+1} \le \delta / s\} / P\{St+1 - St \le \delta / s\}$ | $P\{V_t - V_{t+1} \le \delta / v\} / P\{Vt+1 - Vt \le \delta / v\}$ |
| 1 | 1 | 65 | 3 | 0.771386 | 0.896427 | 0.833872 | 0.847924 | 0.267071 | 0.334452 | . | . | . |
| 2 | 2 | 63 | 3 | 0.730375 | 0.896427 | 0.699981 | 0.682413 | 0.267071 | 0.305153 | 0.800832 | 0.507005 | 0.192826 |
| 3 | 3 | 63 | 3 | 0.730375 | 0.896427 | 0.699981 | 0.682413 | 0.267071 | 0.305153 | 0.748359 | 0.391523 | 0.192826 |
| 4 | 1 | 72 | 2 | 0.811578 | 0.912994 | 0.905528 | 0.909637 | 0.184081 | 0.34027 | 0.499277 | 0.680551 | 0.250393 |
| 5 | 1 | 63 | 2 | 0.730375 | 0.912994 | 0.699981 | 0.669189 | 0.184081 | 0.288159 | 0.426713 | 0.905027 | 0.22607 |
| 6 | 2 | 67 | 4 | 0.771386 | 0.870713 | 0.833872 | 0.853674 | 0.379324 | 0.372728 | 0.800832 | 0.467358 | 0.81271 |
| 7 | 0 | 64 | 1 | 0.730375 | 0.923927 | 0.699981 | 0.655703 | 0.129841 | 0.276449 | 0.510283 | 0.58029 | 0.362439 |
| 8 | 1 | 63 | 3 | 0.730375 | 0.896427 | 0.699981 | 0.682413 | 0.267071 | 0.305153 | 0.838266 | 0.445506 | 0.849854 |

Figure 19: Snapshot of the Nine Probabilities Developed to Test the Validity of an Observation

The observations which had all the nine probabilities less than the threshold specified were identified as valid observations. The observations which had any of the nine probabilities not lying between the thresholds specified were considered invalid. These observations were identified as either partially or totally erroneous, and were further screened to identify the probable erroneous parameters. An observation which had all the three parameters likely to be erroneous was filtered out as totally erroneous observation.

### 6.4 Stage Three: Identification of Erroneous Parameters

A valid observation was used to derive a strategy for identifying the erroneous parameters. Intentional errors were introduced into the valid observations and the patterns among the set of nine probabilities observed by these changes were used to identify the erroneous parameters in general. The screening strategy was devised by fixing one parameter of a valid observation and changing the other parameters to erroneous values. This was based on the assumption that for an observation to be partially valid, at least one of the parameters in the observation should be valid.

Separate analysis was conducted for stable and unstable flow observations as they possessed different probability distributions and would likely possess different patterns when the intentional errors were introduced into the observation. Figure 20 and Figure 21 show snapshots of valid observations (representing stable and unstable flows) that were considered to conduct experiment for identifying the erroneous parameters. An experimental analysis was conducted on each of the above observations to derive the patterns for identifying the erroneous parameters.

| | | | | | Stable flow Condition | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Probabilistic traffic flow relationships | | | | | | Temporal variations | | | |
| O | S | V | P(O≤ok/S=sj) | P(O≤ok/V=vj) | P(S≤si/O=oj) | P(S≤si/V=vj) | P(V≤vi/O=oj) | P(V≤vi/S=sj) | P{Ot-Ot+1≤δ/o} | P{Ot+1-Ot≤δ/o} | P{St-St+1≤δ/s} | P{St+1-St≤δ/s} | P{Vt-Vt+1≤δ/v} | P{Vt+1-Vt≤δ/v} |
| 11 | 60 | 15 | 0.732 | 0.917 | 0.861 | 0.81 | 0.76 | 0.899 | 0.2074 | | 0.62305 | | 0.188 | |

Figure 20: Snapshot of a Valid Observation Representing Stable Flow Condition

| | | | | | Unstable flow Condition | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Probabilistic traffic flow relationships | | | | | | Temporal variations | | | |
| O | S | V | P(O≤ok/S=sj) | P(O≤ok/V=vj) | P(S≤si/O=oj) | P(S≤si/V=vj) | P(V≤vi/O=oj) | P(V≤vi/S=sj) | P{Ot-Ot+1≤δ/o} | P{Ot+1-Ot≤δ/o} | P{St-St+1≤δ/s} | P{St+1-St≤δ/s} | P{Vt-Vt+1≤δ/v} | P{Vt+1-Vt≤δ/v} |
| 28 | 11 | 11 | 0.203 | 0.533 | 0.264 | 0.68 | 0.207 | 0.656 | 0.0802 | | 0.132 | | 0.203 | |

Figure 21: Snapshot of a Valid Observation Representing Unstable Flow Condition

The process was sequentially carried out in eighteen steps to deduce all the patterns which would capture the nature of the most of the erroneous observations with respect to stochastic and conditional variation of the parameters. A valid observation representing stable flow condition was considered first and errors were introduced into the observation. In first six steps, patterns corresponding to single parameter being erroneous were identified while the next twelve steps dealt with identifying the patterns that reflect the invalidity of two parameters. The experimental design to deduce the patterns corresponding to the erroneous parameters is presented in Figure 22.

The probabilistic patterns corresponding to extremely low and high values of parameters were examined separately. This was based upon the reason that the probabilistic nature of the observations differs for the extreme values of the parameters. Intentional errors were first introduced into the valid observation and temporal variation of the parameters was examined to deduce patterns for identifying the erroneous parameters. Validity of an observation was based on how likely is the difference with its preceding observation as mentioned earlier in methodology. Hence, preceding observational values ($O_{t-1}$, $S_{t-1,}$ and $V_{t-1}$) were taken as reference and the absolute difference of occupancy, speed and volume parameters were varied in the experimental sequence designed to correspond to the invalid transitions in a time gap of 30 seconds, implying erroneous nature of the transition. Probabilistic relationship between the parameters was then examined in response to the intentional errors introduced. The derivation of the patterns to identify erroneous parameters (with reference to both temporal variation and probabilistic traffic flow relationship) for stable flow conditions is presented in Figure 23 and Figure 24.

| Pattern | Erroneous parameter | Description |
|---|---|---|
| 1 | Volume- | Only volume parameter is invalid and the value is lower than expected |
| 2 | Volume+ | Only volume parameter is invalid and the value is higher than expected |
| 3 | Speed- | Only speed parameter is invalid and the value is lower than expected |
| 4 | Speed+ | Only speed parameter is invalid and the value is higher than expected |
| 5 | Occupancy- | Only occupancy parameter is invalid and the value is lower than expected |
| 6 | Occupancy+ | Only occupancy parameter is invalid and the value is higher than expected |
| 7 | Speed+, Volume+ | Speed and volume parameters are invalid for a combination of high speed and volume values |
| 8 | Speed+, Volume- | Speed and volume parameters are invalid for a combination of high speed and low volume values |
| 9 | Speed-, Volume+ | Speed and volume parameters are invalid for a combination of low speed and high volume values |
| 10 | Speed-, Volume- | Speed and volume parameters are invalid for a combination of low speed and volume values |
| 11 | Occupancy+, Volume+ | Occupancy and volume parameters are invalid for a combination of high occupancy and high volume values |
| 12 | Occupancy+, Volume- | Occupancy and volume parameters are invalid for a combination of high occupancy and low volume values |
| 13 | Occupancy-, Volume+ | Occupancy and volume parameters are invalid for a combination of low occupancy and high volume values |
| 14 | Occupancy-, Volume- | Occupancy and volume parameters are invalid for a combination of low occupancy and low volume values |
| 15 | Occupancy+, Speed+ | Occupancy and speed parameters are invalid for a combination of high occupancy and speed values |
| 16 | Occupancy+,Speed- | Occupancy and speed parameters are invalid for a combination of high occupancy and low speed values |
| 17 | Occupancy-, Speed+ | Occupancy and speed parameters are invalid for a combination of low occupancy and high speed values |
| 18 | Occupancy-, Speed- | Occupancy and speed parameters are invalid for a combination of low occupancy and lowspeed values |

Figure 22: Snapshot of Patterns Representing Various Erroneous Observations

| Patter no | Erroneous parameter | o | δ | $P\{0_t-0_{t+1}\leq\delta/o\}/P\{0t+1-0t\leq\delta/o\}$ | s | δ | $P\{S_t-S_{t+1}\leq\delta/s\}/P\{St+1-St\leq\delta/s\}$ | v | δ | $P\{V_t-V_{t+1}\leq\delta/v\}/P\{Vt+1-Vt\leq\delta/v\}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | Default | 10 | 1 | 0.480 | 54 | 6 | 0.623 | 14 | 1 | 0.460 |
| 1 | Volume- | 10 | 1 | 0.480 | 54 | 6 | 0.623 | 14 | 13 | 0.923 |
| 2 | Volume+ | 10 | 1 | 0.480 | 54 | 6 | 0.623 | 14 | 6 | 0.960 |
| 3 | Speed- | 10 | 1 | 0.480 | 54 | 54 | 0.992 | 14 | 1 | 0.460 |
| 4 | Speed+ | 10 | 1 | 0.480 | 54 | 46 | 1.000 | 14 | 1 | 0.460 |
| 5 | Occupancy- | 10 | 10 | 0.910 | 54 | 6 | 0.623 | 14 | 1 | 0.460 |
| 6 | Occupancy+ | 10 | 90 | 0.972 | 54 | 6 | 0.623 | 14 | 1 | 0.460 |
| 7 | Speed+, Volume+ | 10 | 1 | 0.480 | 54 | 46 | 1.000 | 14 | 6 | 0.960 |
| 8 | Speed+, Volume- | 10 | 1 | 0.480 | 54 | 46 | 1.000 | 14 | 13 | 0.923 |
| 9 | Speed-, Volume+ | 10 | 1 | 0.480 | 54 | 54 | 0.992 | 14 | 6 | 0.960 |
| 10 | Speed-, Volume- | 10 | 1 | 0.480 | 54 | 54 | 0.992 | 14 | 13 | 0.923 |
| 11 | Occupancy+, volume+ | 10 | 90 | 0.972 | 54 | 6 | 0.623 | 14 | 6 | 0.960 |
| 12 | Occupancy+, volume- | 10 | 90 | 0.972 | 54 | 6 | 0.623 | 14 | 13 | 0.923 |
| 13 | Occupancy-, volume+ | 10 | 10 | 0.910 | 54 | 6 | 0.623 | 14 | 6 | 0.960 |
| 14 | Occupancy-, volume- | 10 | 10 | 0.910 | 54 | 6 | 0.623 | 14 | 13 | 0.923 |
| 15 | Occupancy+, speed+ | 10 | 90 | 0.972 | 54 | 46 | 1.000 | 14 | 1 | 0.460 |
| 16 | Occupancy+,speed- | 10 | 90 | 0.972 | 54 | 54 | 0.992 | 14 | 1 | 0.460 |
| 17 | Occupancy-, speed+ | 10 | 10 | 0.910 | 54 | 46 | 1.000 | 14 | 1 | 0.460 |
| 18 | Occupancy-, speed- | 10 | 10 | 0.910 | 54 | 54 | 1.000 | 14 | 1 | 0.460 |

Figure 23: Derivation of Patterns Representing Various Erroneous Observations in Stable Flow Conditions (With Respect to Approach One)

| pattern no | Erroneous parmeter | O | S | V | P(O≤ok/S=sj) | P(O≤ok/V=vj) | P(S≤si/O=oj) | P(S≤si/V=vj) | P(V≤vi/O=o | P(V≤vi/S=sj) |
|---|---|---|---|---|---|---|---|---|---|---|
| default | | 2 | 12 | 15 | 0.732 | 0.916 | 0.86 | 0.809 | 0.759 | 0.899 |
| 1 | Volume - | 2 | 12 | 1 | 0.732 | 0.951 | 0.86 | 0.657 | 0.03 | 0.248 |
| 2 | Volume + | 2 | 12 | 20 | 0.732 | 0.644 | 0.86 | 0.841 | 0.92 | 0.923 |
| 3 | Speed - | 2 | 0 | 15 | 0.748 | 0.916 | 0.077 | 0.368 | 0.759 | 0.998 |
| 4 | Speed + | 2 | 20 | 15 | 0.766 | 0.916 | 0.97 | 0.969 | 0.759 | 0.942 |
| 5 | Occupancy - | 0 | 12 | 15 | 0.733 | 0.173 | 0.694 | 0.785 | 0.903 | 0.809 |
| 6 | Occupancy + | 20 | 12 | 15 | 1 | 1 | 0.992 | 0.785 | 0.905 | 0.809 |
| 7 | Speed+, Volume+ | 2 | 20 | 20 | 0.766 | 0.664 | 0.97 | 0.969 | 0.923 | 0.961 |
| 8 | Speed+, Volume- | 2 | 20 | 0 | 0.791 | 0.951 | 0.97 | 0.966 | 0.03 | 0.433 |
| 9 | Speed-, Volume+ | 2 | 0 | 20 | 0.791 | 0.664 | 0.077 | 0.342 | 0.914 | 1 |
| 10 | Speed-, Volume- | 2 | 0 | 0 | 1 | 0.951 | 0.077 | 0.947 | 0.33 | 0.132 |
| 11 | Occupancy+, Volume+ | 20 | 12 | 20 | 1 | 1 | 0.992 | 0.84 | 0.914 | 0.928 |
| 12 | Occupancy+, Volume- | 20 | 12 | 0 | 0.733 | 0.78 | 0.992 | 0.657 | 0.33 | 0.248 |
| 13 | Occupancy-, Volume+ | 0 | 12 | 20 | 0.733 | 0.128 | 0.694 | 0.841 | 0.914 | 0.928 |
| 14 | Occupancy-, Volume- | 0 | 12 | 0 | 1 | 0.901 | 0.694 | 0.65 | 0.198 | 0.248 |
| 15 | Occupancy+, Speed+ | 20 | 20 | 15 | 0.99 | 1 | 0.97 | 0.965 | 0.905 | 0.942 |
| 16 | Occupancy+, Speed- | 20 | 20 | 15 | 0.631 | 1 | 0.947 | 0.368 | 0.905 | 0.942 |
| 17 | Occupancy-, Speed+ | 0 | 20 | 15 | 0.67 | 0.714 | 0.951 | 0.969 | 0.903 | 0.942 |
| 18 | Occupancy-, Speed- | 0 | 0 | 15 | 0.698 | 0.714 | 0.08 | 0.368 | 0.903 | 0.998 |

Figure 24: Derivation of Patterns Representing Various Erroneous Observations in Stable Flow Conditions (With Respect to Approach Two)

Similar experimental analysis was conducted on the unstable flow observation and the patterns corresponding to the changes made with respect to stochastic and conditional variation of the parameters were identified. The changes made in accordance with the experimental design and the corresponding probabilistic patterns observed due to these changes are presented in Figure 25 and Figure 26.

### 6.5 Results and Interpretation of Stage Three

This section presents the results obtained from the six case studies conducted on the stable and unstable flow observations and the process of screening the observation with reference to the patterns derived. Figure 27 represents the 18 patterns that were derived form the experimental analysis conducted on a valid stable flow observation considered. From Figure 27 it can be seen that pattern 5 doesn't indicate the erroneous nature of the observation for low values of occupancy. A possible explanation to this would be that the occupancy values for the stable flow conditions are quiet low (0-20%) and a drop from these values is feasible for all the combinations of valid speed and volume parameters.

Similarity among the patterns 2 and 13, 3 and 18, 4 and 17 as shown in the Figure 27 could be attributed to the fact that maximum occupancy drop for stable flow (i.e. for low occupancy range conditions is feasible as explained earlier and doesn't have any effect by itself when combined with other erroneous parameters. The patterns developed could be used to identify the erroneous parameters. For instance, if an observation has probabilities matching with the pattern 1 (volume-), a conclusion that the volume parameter is erroneous and the value is less than expected to be valid is reached.

| Patter no | Erroneous parameter | o | δ | $P\{0_t-0_{t+l}\leq\delta/o\}/P\{0t+1-0t\leq\delta/o\}$ | s | δ | $P\{S_t-S_{t+l}\leq\delta/s\}/P\{St+1-St\leq\delta/s\}$ | v | δ | $P\{V_t-V_{t+l}\leq\delta/v\}/P\{Vt+1-Vt\leq\delta/v\}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | Default | 26 | 2 | 0.080 | 10 | 1 | 0.133 | 11 | 0 | 0.204 |
| 19 | volume - | 26 | 2 | 0.080 | 10 | 1 | 0.133 | 11 | 10 | 0.904 |
| 20 | volume + | 26 | 2 | 0.080 | 10 | 1 | 0.133 | 11 | 9 | 0.988 |
| 21 | speed - | 26 | 2 | 0.080 | 10 | 10 | 1.007 | 11 | 0 | 0.204 |
| 22 | speed + | 26 | 2 | 0.080 | 10 | 90 | 1.024 | 11 | 0 | 0.204 |
| 23 | occupancy - | 26 | 26 | 0.984 | 10 | 1 | 0.132 | 11 | 0 | 0.203 |
| 24 | occupancy + | 26 | 74 | 0.991 | 10 | 1 | 0.132 | 11 | 0 | 0.203 |
| 25 | speed+, volume+ | 26 | 2 | 0.080 | 10 | 90 | 1.024 | 11 | 9 | 0.988 |
| 26 | speed+, volume- | 26 | 2 | 0.080 | 10 | 90 | 1.024 | 11 | 10 | 0.904 |
| 27 | speed-, volume+ | 26 | 2 | 0.080 | 10 | 10 | 1.007 | 11 | 9 | 0.988 |
| 28 | speed-, volume- | 26 | 2 | 0.080 | 10 | 10 | 1.007 | 11 | 10 | 0.904 |
| 29 | occupancy+, volume+ | 26 | 74 | 0.991 | 10 | 1 | 0.132 | 11 | 9 | 0.988 |
| 30 | occupancy+, volume- | 26 | 74 | 0.991 | 10 | 1 | 0.132 | 11 | 10 | 0.905 |
| 31 | occupancy-, volume+ | 26 | 26 | 0.984 | 10 | 1 | 0.132 | 11 | 9 | 0.988 |
| 32 | occupancy-, volume- | 26 | 26 | 0.984 | 10 | 1 | 0.132 | 11 | 10 | 0.905 |
| 33 | occupancy+, speed+ | 26 | 74 | 0.991 | 10 | 90 | 1.020 | 14 | 1 | 0.189 |
| 34 | occupancy+, speed- | 26 | 74 | 0.991 | 10 | 10 | 1.006 | 14 | 1 | 0.189 |
| 35 | occupancy-, speed+ | 26 | 26 | 0.984 | 10 | 90 | 1.020 | 14 | 1 | 0.189 |
| 36 | occupancy-, speed- | 26 | 26 | 0.984 | 10 | 10 | 1.006 | 14 | 1 | 0.189 |

Figure 25: Derivation of Patterns Representing Various Erroneous Observations in Unstable Flow Conditions (With Respect to Approach One)

| pattern no | Erroneous parmeter | O | S | V | $P(O \le o_k / S = s_j)$ | $P(O \le o_k / V = v_j)$ | $P(S \le s_i / O = o_j)$ | $P(S \le s_i / V = v_j)$ | $P(V \le v_i / O = o_j)$ | $P(V \le v_i / S = s_j)$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 19 | Volume - | 5 | 2 | 1 | 0.2030384 | 0.085322 | 0.263805 | 0.97011 | 0.104881 | 0.0986 |
| 20 | Volume + | 5 | 2 | 20 | 0.2030384 | 0.8738 | 0.263805 | 0.3617 | 0.846 | 1 |
| 21 | Speed - | 5 | 0 | 11 | 0.872814 | 0.533433 | 0.198 | 0.4351 | 0.2069 | 0.91605 |
| 22 | Speed + | 5 | 20 | 11 | 0.957803 | 0.533433 | 0.978 | 0.972 | 0.2069 | 0.878 |
| 23 | Occupancy - | 0 | 2 | 11 | 0.203 | 0.3146 | 0.099 | 0.679 | 0.8104 | 0.6561 |
| 24 | Occupancy + | 20 | 2 | 11 | 0.994 | 1 | 0.9626 | 0.679 | 0.8727 | 0.6561 |
| 25 | Speed+, Volume+ | 5 | 20 | 20 | 0.957803 | 0.8738 | 0.978 | 0.969 | 0.846 | 0.96013 |
| 26 | Speed+, Volume- | 5 | 20 | 1 | 0.957803 | 0.085322 | 0.978 | 0.968 | 0.104881 | 0.433 |
| 27 | Speed-, Volume+ | 5 | 0 | 20 | 0.872814 | 0.8738 | 0.198 | 0.3424 | 0.846 | 1 |
| 28 | Speed-, Volume- | 5 | 0 | 1 | 0.872814 | 0.085322 | 0.198 | 0.9471 | 0.104881 | 0.1328 |
| 29 | Occupancy+, Volume+ | 20 | 2 | 20 | 0.994 | 1 | 0.9626 | 0.3617 | 0.9141 | 1 |
| 30 | Occupancy+, Volume- | 20 | 2 | 1 | 0.994 | 0.784 | 0.9626 | 0.97011 | 0.333 | 0.098 |
| 31 | Occupancy-, Volume+ | 0 | 2 | 20 | 0.203 | 0.128 | 0.099 | 0.3617 | 0.941 | 1 |
| 32 | Occupancy-, Volume- | 0 | 2 | 1 | 0.203 | 0.901 | 0.099 | 0.97011 | 0.198 | 0.098 |
| 33 | Occupancy+, Speed+ | 20 | 20 | 11 | 1 | 1 | 0.998 | 0.97 | 0.872 | 0.878 |
| 34 | Occupancy+, Speed- | 20 | 0 | 11 | 0.998 | 1 | 0.9466 | 0.453 | 0.872 | 0.916 |
| 35 | Occupancy-, Speed+ | 0 | 20 | 11 | 0.67 | 0.314 | 0.96 | 0.97 | 0.8104 | 0.878 |
| 36 | Occupancy-, Speed- | 0 | 0 | 11 | 0.698 | 0.314 | 0.087 | 0.453 | 0.8104 | 0.916 |

Figure 26: Derivation of Patterns Representing Various Erroneous Observations in Unstable Flow Conditions (With Respect to Approach Two)

| Pattern No | Erroneous parameter | $P(O \le o_k / S = s_j)$ | $P(O \le o_k / V = v_j)$ | $P(S \le s_j / O = o_j)$ | $P(S \le s_j / V = v_j)$ | $P(V \le v_j / O = o_j)$ | $P(V \le v_j / S = s_j)$ | $P\{O_t - O_{t-1} \le \delta/o\} / P\{Ot+1 - Ot \le \delta/o\}$ | $P\{S_t - S_{t-1} \le \delta/s\} / P\{St+1 - St \le \delta/s\}$ | $P\{V_t - V_{t-1} \le \delta/v\} / P\{Vt+1 - Vt \le \delta/v\}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Volume- | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | Volume+ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 3 | Speed- | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 4 | Speed+ | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| 5 | Occupancy- | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | Occupancy+ | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| 7 | Speed+, Volume+ | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| 8 | Speed+, Volume- | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| 9 | Speed-, Volume+ | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 |
| 10 | Speed-, Volume- | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 11 | occupancy+, volume+ | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| 12 | occupancy+, volume- | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| 13 | occupancy-, volume+ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 14 | occupancy-, volume- | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 15 | occupancy+, speed+ | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| 16 | occupancy+,speed- | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| 17 | occupancy-, speed+ | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| 18 | occupancy-, speed- | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |

[3]

Figure 27: Snapshot of Patterns for Screening the Stable Flow Observations

[3] 1- Cumulative probability >.95, 0-cumulative probability <.95.

Incase an observation matches with overlapping patterns (for example 2 and 13), it could be concluded that either occupancy or volume parameter is erroneous but a definitive conclusion that both of the parameters are erroneous could not be reached.

The patterns derived from the experimental analysis conducted on the unstable flow observation considered are presented in the Figure 28. Similarity between patterns 28 and 21 i.e. the patterns representing speed- and volume-, and speed- conditions was observed. This could be attributed to the fact that the low value of the volume parameter considered doesn't have an effect on the erroneous nature of the observation when combined with low values of speed. Patterns 22 and 26 overlap suggesting that either speed or volume parameter is erroneous.

The real-time data set was first screened with a threshold value of .95 as mentioned earlier. The erroneous observations (with probabilities greater than .95) were further matched with the patterns developed to identify erroneous parameters. The implementation of data screening algorithm for the real-time data is presented in Figure 29. Table 4 shows the results of screening process conducted. The results from the table showed that 75% of the observations were likely to be valid, while the remaining 20% observations were identified as partially valid observations. The remaining 5% of the observations could be either totally erroneous observations or observations representing conflicting conclusion. These conflicting conclusions could be reached when an observation matches with two or more patterns which indicate contradictory results.

| Pattern No | Condition | $P(O \leq o_k / S = s_j)$ | $P(O \leq o_k / V = v_j)$ | $P(S \leq s_j / O = o_j)$ | $P(S \leq s_j / V = v_j)$ | $P(V \leq v_j / O = o_j)$ | $P(V \leq v_j / S = s_j)$ | $P\{O_t - O_{t+1} \leq \delta / o\} / P\{Ot+1 - Ot \leq \delta / o\}$ | $P\{S_t - S_{t+1} \leq \delta / s\} / P\{St+1 - St \leq \delta / s\}$ | $P\{V_t - V_{t+1} \leq \delta / v\} / P\{Vt+1 - Vt \leq \delta / v\}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 19 | Volume- | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 20 | Volume+ | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 21 | Speed- | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 22 | Speed+ | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| 23 | Occupancy- | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 24 | Occupancy+ | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| 25 | Speed+, Volume+ | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| 26 | Speed+, Volume- | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| 27 | Speed-, Volume+ | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 |
| 28 | Speed-, Volume- | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 29 | occupancy+, volume+ | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 |
| 30 | occupancy+, volume- | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 |
| 31 | occupancy-, volume+ | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |
| 32 | occupancy-, volume- | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| 33 | occupancy+, speed+ | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| 34 | occupancy+,speed- | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| 35 | occupancy-, speed+ | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| 36 | occupancy-, speed- | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |

Figure 28: Snapshot of Patterns for Screening the Unstable Flow Observations

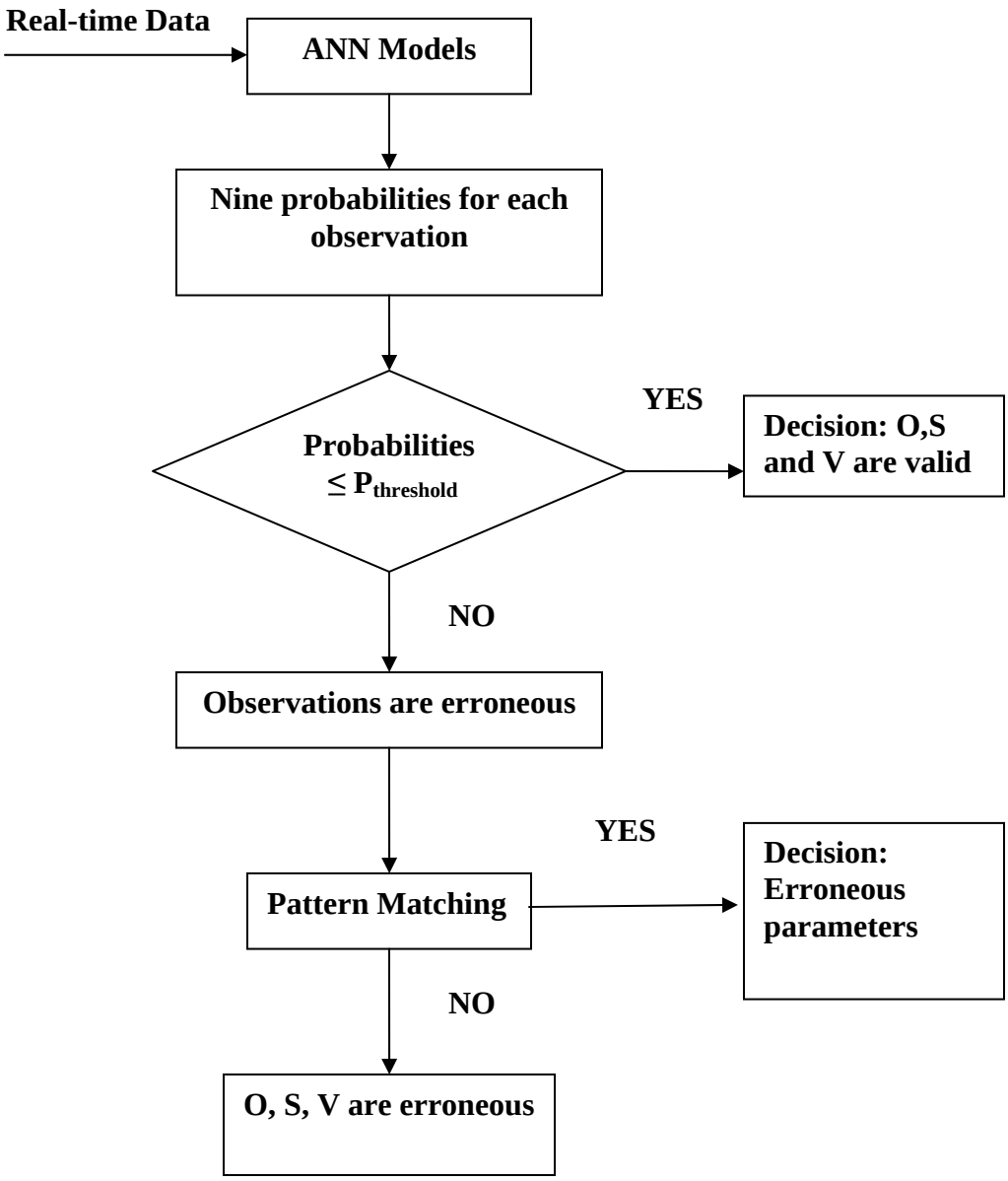


Figure 29: Implementation of Data Screening Algorithm for Real-time Traffic Data

Table 4: Results of Implementation of Data screening Algorithm on Real-time Data

| Pattern No | Erroneous parameters | Total number of observations- 51837 | | Percentage of observations |
|---|---|---|---|---|
| | | Valid observations | Invalid observations | |
| | | 38657 | - | 74.57 |
| 1 | Volume- | - | 3997 | 7.71 |
| 2 | Volume + | - | 417 | 0.804 |
| 3 | Speed - | - | 23 | 0.044 |
| 4 | Speed + | - | 76 | 0.146 |
| 7 | Speed +, Volume + | - | 7 | 0.013 |
| 8 | Speed +, Volume - | - | 19 | 0.036 |
| 9 | Speed- Volume + | - | 40 | 0.077 |
| 15 | Occupancy +, speed+ | - | 3745 | 7.22 |
| 16 | Occupancy +, speed - | - | 2043 | 3.94 |
| 18 | Occupancy -, speed - | - | 56 | 0.108 |
| 21 | Speed- (unstable flow) | - | 19 | 0.036 |
| 22 | Speed + ( unstable flow) | - | 9 | 0.017 |
| 23 | Occupancy- (unstable flow) | - | 114 | 0.219 |
| 24 | Occupancy + (unstable flow) | - | 9 | 0.017 |
| 25 | Speed +, volume+ (unstable flow) | - | 113 | 0.217 |
| 27 | Speed -, volume+ (unstable flow) | - | 7 | 0.013 |

# 7. SUMMARY AND CONCLUSIONS

## 7.1 Study Summary

The research study presented a probabilistic approach for real-time screening of freeway loop detector data. A stream of continuous data, compiled from inductive loop detectors over the 38 mile corridor of I-4 Orlando, Florida, was used in this study. A methodology for screening observations was formulated on the basis of feasible stochastic and probabilistic variations of the three macroscopic traffic parameters (Speed, Occupancy and Volume).

Experimental work was carried out in three stages. In the first stage, detector data complied from the year 2000 was used to examine the temporal variation of each of the three traffic parameters to determine the likelihood of observing abrupt changes in traffic conditions over time. Probabilistic traffic flow relationships were used to check for inconsistencies between the three parameters. These approaches were complementary to each other in the sense that an observation was validated by using both, feasible temporal variations, and consistency of the relationship between the three traffic parameters. These distribution functions were approximated in the second stage using Multi-layer Feed-forward Networks. A total of 38 networks were developed to model the probabilities representing these variations efficiently. The trained networks were then used to produce the probabilities associated with a twenty-four hour real-time data set that was extracted in the year 2002 from 70 dual loop detector stations. In the third stage a demonstration of the logic contemplated to devise a screening algorithm was presented. The probabilities derived from the networks were then screened using a 95% threshold to filter out the erroneous observations. The erroneous observations were further matched with the

patterns that were generated to identify the erroneous parameters that contribute to the invalidity of the observation.

### 7. 2 Conclusions

The non-linear nature of the stochastic and conditional relationships between the parameters was reasonably captured using 38 Multi-layer Feed-forward Networks, except for the stochastic variations representing high occupancy conditions. The screening algorithm devised was efficient in judging the validity of the real-time data format with 95% probability. Most of the patterns deduced were capable of identifying the erroneous parameters in the observation thus classifying them into partially valid observations. The following were contributions of this research study.

This approach can be implemented online or offline to screen the observations before or after they are streamed into data warehouses and is user adaptable as it does not impose any restrictions on the thresholds. This approach could be used for imputing the erroneous parameters in the sense that the patterns deduced were capable of identifying the erroneous parameter, and also communicated information about the dimension of the erroneous parameter (for example volume- suggest that volume parameter is erroneous and the value is less than expected). Thus this study provides preliminary information required for imputation of erroneous parameters. The general approach formulated in this study can be applied to different locations while the transferability of the model needs to be authenticated with more tests using the data from other locations and are not dealt with in this study. The algorithm can also be used to identify the operational status of detectors and detect calibration problems that may call for immediate maintenance due to its real-time screening ability.

### 7.3 Limitations and Future Research

The study conducted has the following limitations:

- The model is only applicable for continuous data as the temporal variation of the parameters approach uses two successive observations with a time gap of 30 seconds. An observation compares itself with the preceding observation and checks feasibility of the temporal variation. If the preceding observation is itself erroneous then approach 1 is not applicable. Such observations could still be validated by considering Probabilistic traffic flow relationship approach but the research study loses its grip in identifying partially valid observations with two erroneous parameters accurately.
- This approach could not explain the cases where an observation matches with two or more patterns leading to conflicting conclusions. For example, if an observation matches with both (speed –) and (speed +, volume-), it shall lead to conflicting conclusions that speed parameter is erroneous and is both high and low. These cases could not be explained by the approach and needs to be researched further.
- The model developed is applicable for data obtained under same operating conditions. In simple terms, the model uses historical data and assumes that same operating conditions exist in the future as well. This might not be true in cases such as adding up of lanes etc. Thus the model needs recalibration whenever different operating conditions exist.

Future research may be conducted to attend to these limitations and also to check the issues regarding the transferability of the model to different locations. This attempt if

successful can impact the online detection maintenance system through out the country.

# REFERENCES


1. Chen, L., and May A. Traffic Detector Errors and Diagnostics. In *Transportation Research Record 1132*, TRB, National Research Council, Washington D.C., 1987, pp. 82-93.

2. Coifman, B. Using Dual Loop Speed Traps to Identify Detector Errors. In *Transportation Research Record 1683*, TRB, National Research Council Washington D.C., 1999, pp. 47-58.

3. Coifman, B. and S. Dhoorjaty. Event Data Based Traffic Detector Validation Tests. Presented at 81st Annual Meeting of the Transportation Research Board, Washington D.C., 2002.

4. Jacobson, L.N., N.L. Nihan, and J.D. Bender. Detecting Erroneous Loop Detector Data in a Freeway Traffic Management System. In Transportation Research Record 1287, TRB, National Research Council, Washington D.C., 1990, pp. 151-166.

5. Cleghorn, D., F. L. Hall, and D. Garbuio. Improved Data Screening Techniques for Freeway Traffic Management Systems. In Transportation Research Record *1320*, TRB, National Research Council, Washington D.C., 1991, pp. 17-23.

6. H J. Payne AND S.Thompson. Malfunction Detection and Data Repair for Induction-Loop Sensors Using I-880 Data Base. Transportation Research Record 1570. Paper No. 971113191. pp 191-201.

7. Turochy, R.E. and B. Smith. A New Procedure for Detector Data Screening in Traffic Management Systems. In Transportation Research Record 1727, TRB, National Research Council, Washington D.C., 2000, pp. 127-131.

8. Peeta, S., and I. Anastassopoulos. Automatic Real-Time Detection and Correction of Erroneous Detector Data Using Fourier Transforms for On-Line Traffic Control Architectures. Presented at 81st Annual Meeting of the Transportation Research Board, Washington D.C., 2002.

9. Ishak, S. Quantifying uncertainties of freeway detector Observations using fuzzy clustering approach. TRB 03-2056, Transportation Research Board 82nd Annual Meeting Washington, D. C. 2003.

10. C. Chen. Detecting Errors and Imputing Missing Data for Single Loop Surveillance Systems. TRB 03-3507, 82nd Annual Meeting Transportation Research Board January 2003 Washington, D.C.

11. Z. Wall and D.J. Dailey. An Algorithm for the Detection and Correction of Errors in Archived Traffic Data. TRB 03-4184, Annual Meeting Transportation Research Board Washington, D.C.2003.

12. Chilakamarri V.S.R.C, H M. Al-Deek, Revised New Algorithms for Filtering and Imputation of Real Time and Archived Dual-Loop Detector Data. TRB 2004 annual meeting. Paper No. 04-3505.

13. Nihan, N. Aid to Determining Freeway Metering Rates and Detecting Loop Errors. Journal of Transportation Engineering, ASCE, Vol. 123, No 6, November/December 1997, pp. 454-458.

14. Wang, Y. and N. Nihan. Freeway Traffic Speed Estimation Using Single Loop Outputs. In Transportation Research Record 1727, TRB, National Research Council, Washington D.C., 2000, pp. 120-126.

15. Wang, Y. and Nihan. A Robust Method of Filtering Single-Loop Data for Improved Speed Estimation. Presented at 81st Annual Meeting of the Transportation Research Board, Washington D.C., 2002.

16. Roess, R.P., W. R. McShane, and E. S. Prassas. Traffic Engineering, second edition, Prentice Hall, New Jersey, 1998.

17. Walpole, R.E., Myers, Y.E. Probability & Statistics for engineers and scientists seventh edition, Prentice Hall, New Jersey, 2002.

18. R. J.Freund., W. J.Wilson. Statistical Methods, Academic Press, London 1996.

19. S. Haykin. Neural networks, Macmillan publications, New York 1994.

# VITA

Shourie Kondagari was born on April 9, 1982, in city of Hyderabad, Andhra Pradesh, India. He obtained his Bachelor of Engineering degree in Civil Engineering in 2003 from Osmania University, Hyderabad, India. He joined Louisiana State University in August 2004 for his Masters pursuit and expects to receive the degree of Master of Science in Civil Engineering in August 2006.